Binding determinants of High Mobility Group proteins in the mouse genome

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Daniele Filippo Maria Colombo

von Italien

Basel, 2017

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von Prof. Dr. Dirk Schübeler, Prof. Dr. Bart Deplancke

Basel, den 15-11-2016

Prof. Dr. Jörg Schibler
Dekan der Fakultät

PhD Thesis | 2016

Binding determinants of High Mobility Group proteins in the mouse genome

Daniele Filippo Colombo

To Christiane Wirbelauer

Who suffers now and who taught me the most important teaching of my PhD

1 Index

2 Frequent abbreviations

aa: amino acid

AT-rich: W nucleotides rich

bioChIP: biotin mediated ChIP

ChIP: cromatin immunoprecipitation

DBD: DNA binding domain

DNA: deoxyribonucleic acid

Fox: Forkhead box

GFP: Green fluorescent protein

HMG: high mobility group

KO: knock out

MBD: methyl-binding domain

MCS: minimal cloning site

Mm: Mus musculus

NGS: next generation sequencing

RAMBiO: recombinase-assisted mapping of biotin-tagged proteins (Baubec et al., 2013)

RNA: ribonucleic acid

SAV: streptavidin

SELEX: Systematic evolution of ligands by exponential enrichment

seq: sequencing

TF: Transctiption factor

KO: knock out

WT: wild type

PTM: post translational modification

3 Summary

In this work we investigate the determinants of recruitment to DNA and chromatin for HMGB1-2-3-4 and HMGA1-2 and a selection of transcription factors (TF). We adopt a mouse embryonic stem cell (ESC) model system for the generation of antibody independent ChIP-sequencing data. We first report successful recapitulation of Sox2 binding, our internal control, and then focus on HMGA and HMGB proteins, for which no exhaustive genome-wide data had been available.

In the nucleus HMG proteins are one of the major chromatin-associated non-histone proteins. As such they have been implicated in a wide range of nuclear processes from transcription, to nucleosome remodeling, DNA damage and apoptosis.

For HMGB proteins we show frequent contacts with active regulatory regions, which however are also sites of preferred interaction for sequence-unspecific DNA binders and inert proteins such as DNAseI or monomeric GFP. Upon mutation of the DNA binding domains of Hmgb1 no change in the localization pattern for this protein is observed. Additionally upon Hmgb1 knock out (KO), ESC do not show alterations in transcription, as one would expect for a protein involved in regulatory functions. Nevertheless we cannot formally exclude that the biotin tagging is causing a mislocalization of the HMGB proteins, nor that upon Hmgb1 KO HMGB2 may compensate for HMGB1 absence.

As far as HMGA1 and HMGA2 are concerned, on the contrary we show binding throughout the genome with a preference for AT-rich DNA. Mutation of key residues in the DNA binding domains of both proteins causes loss of the AT dependence and the residual signal is comparable to that of a freely diffusing protein (monomeric GFP). Importantly AT-rich dependence is independent of chromatin states, as exemplified by invariance upon neuronal differentiation. These results highlight the fact that the three DNA binding domains of HMGA1 and HMGA2 are the sole determinants of their genomic distribution.

At the chromosomal scale, we show that enriched regions are also generally positive for features of heterochromatin such as presence of Histone H3 Lysine9 methylation, their late replication in S phase and their association with the nuclear lamina. This data points to enrichment of HMGA1-2 at constitutive

heterochromatin, which has a known compositional bias. Lastly, we show a limited role for HMGA1 in the regulation of transcription in ESC by profiling expression patterns of an isogenic KO cell line.

Taken together, the findings on HMGA proteins reveal a broad DNA-binding modality, which supports their known preference for AT-rich DNA. At the same time, our genomic and gene expression results are in contrasts with the often-mentioned roles in transcriptional regulation.

4 Introduction

4.1 Foreword

Strong is the fascination still for developmental scientist, on how from a single fertilized egg life achieves the dazzling cellular specialization observed in multicellular organisms.

As nuclear transplantation experiment have shown the secret lies in the nucleic acid, or better the genome, but also requires specific gametal/early embryonic protein factors (Kang et al., 2014). This is similar to reading a novel: in order to connect all the different episodes that are narrated, it has to be read from the beginning. However how specific trajectories of genome readout are taken molecularly, maintained or reversed is not yet fully understood, thus the fascination remains.

In recent years we have started to collect valuable in vivo information on how individual parts of the nuclear system are working at specific time points. Nevertheless, our inability so far to come up with predictive models for cell fate homeostasis and transition highlights that either we haven't managed to put all pieces in the correct place or that we are still missing relevant information for some overlooked or poorly studied components.

With this thesis I am summarizing our findings for High mobility group proteins, a class of DNA-binders whose in vivo binding properties have been poorly characterized. We are convinced that this data will contribute to make current models of genome biology more precise.

To put our work in perspective, in the introductory section below, I am first resuming what the genome biology community knows about the other, better-studied actors. After that, and before moving to our own results, I will sum up existing biological evidence that links High mobility group proteins to nuclear biology.

4.1.1 The basis of cell identity: a complex code for the complexity of life

We know that cell fate homeostasis and transitions are mainly dictated by gene expression (Moris et al., 2016). It follows that a molecular understanding of expression control would lead to better models for cell fate predictions.

One fundamental form of gene control is regulation of transcription. Transcription in metazoans is generally regulated through the concerted activity of DNA elements called promoters and enhancers (Levine, 2010).

Promoters are those regions of the genome where transcription of a gene starts. In unicellular organisms, promoters tend to contain all the elements for correct assembly of a productive RNA-polymerase in basal conditions or upon response to stimuli. Controlling gene expression at the promoter level however is very ineffective in multicellular organism. Different cellular identity typically results from changes in the spatiotemporal regulation of gene expression during development (Wray, 2007). The control mechanism that allows differential gene expression with fewer or no pleiotropic effects relies on placing the regulatory regions far from gene promoters (Wittkopp and Kalay, 2011). Thus, mutations, which are the drivers of evolution, in one regulatory region will not affect the function of the protein itself or of regulatory regions associated with the same gene in different cell types (Carroll, 2008).

It thus appears that key for multicellularity was a change in the modality of gene regulation, rather than gene innovation (Sebé-Pedrós et al., 2016). In other words, distal regulatory regions (also called enhancers) are by-products of multicellularity. Proof for this is the fact that the unicellular ancestor of animals already had the complex repertoire of genes linked to multicellular processes (de Mendoza et al., 2013).

However with this invention also came the problem of how to connect faithfully promoters with enhancer regions. The solution was probably in the repurposing of a TF, from controlling gene expression, to allowing contacts between regulatory regions. In mammals we known that CTCF and Cohesin perform this fundamental function (Ing-Simmons et al., 2015), and indeed Ctcf KO is lethal at the pre-implantation stage (Ong and Corces, 2014).

A recent study describing the effect of inverting a single CTCF site highlighted how difficult it is to target the proper promoter to the correct enhancer in the

dense nuclear environment (Guo et al., 2015). This notion is also supported by the fact that genes that need less of temporal and spatial control of expression (constitutive or exquisitely cell type specific) tend to maintain the majority of information encoded in the promoter regions (Carroll, 2008; Heidari et al., 2014).

However the reality is much more complex, for example we know now that the frequency of enhancer-promoter contacts is modulated by the general transcriptional state of a region (Whalen et al., 2016) and by larger three dimensional structures called topologically associated domains (TAD)(Ciabrelli and Cavalli, 2015). Additionally, the binding of a single TF usually does not activate an enhancer. Oftentimes groups of TFBSs function together to direct gene expression from a specific enhancer (Yáñez-Cuna et al., 2013). The combinatorial nature of these groupings gives enhancers the ability to integrate inputs from multiple TFs, in order to direct the spatial and temporal patterns of gene expression in very complex ways (Andersson et al., 2014).

This complexity with which enhancer-promoter interactions achieve transcriptional control also opens opportunities for fast speciation, in a delicate balance between conservation and innovation (Prescott et al., 2015; Villar et al., 2015).

4.2 DNA binding in the context of chromatin

As demonstrated above, interaction between DNA and DNA-binding proteins is essential for gene transcription. Upon binding to such regulatory regions these proteins, which are called transcription factors (TF), can initiate transcription, fine-tune it or repress it. There are classes of proteins that have DNA-binding sites embedded in enhancer or promoter sequences.

However it is still unclear why in vivo TFs only bind to a minority of the DNA sequences that have similar nucleotide composition to the preferred binding sites as determined in vitro (Slattery et al., 2014). This uncertainty is one of the major hurdles that we need to overcome if we want to correctly predict transcriptional states.

Recently a lot of effort has been put to address this issue of DNA binding control. The picture that is emerging is that many different mechanisms are at work at

the same time and that each DNA binder obeys to a different set of rules (Spitz and Furlong, 2012).

The most common layers of binding control include TF binding sites clustering, context dependence and DNA shape, nucleosomal occupancy, DNA topology and finally epigenetic, both on the DNA substrate (DNA methylation) and at the chromatin level (via histone PTM/variants and the associated protein complexes modulating accessibility). This fine balance is fundamental for cell differentiation and homeostasis, and when the balance it´s broken, deregulated cells may cause important diseases such as cancer (Shah et al., 2014; 2013).

Since we wanted to investigate the binding determinants of High mobility group proteins (Figure 4-1), here below I will adopt more a protein-centered perspective and for each class of well-characterized DNA binder I will highlight their major determinants of binding.

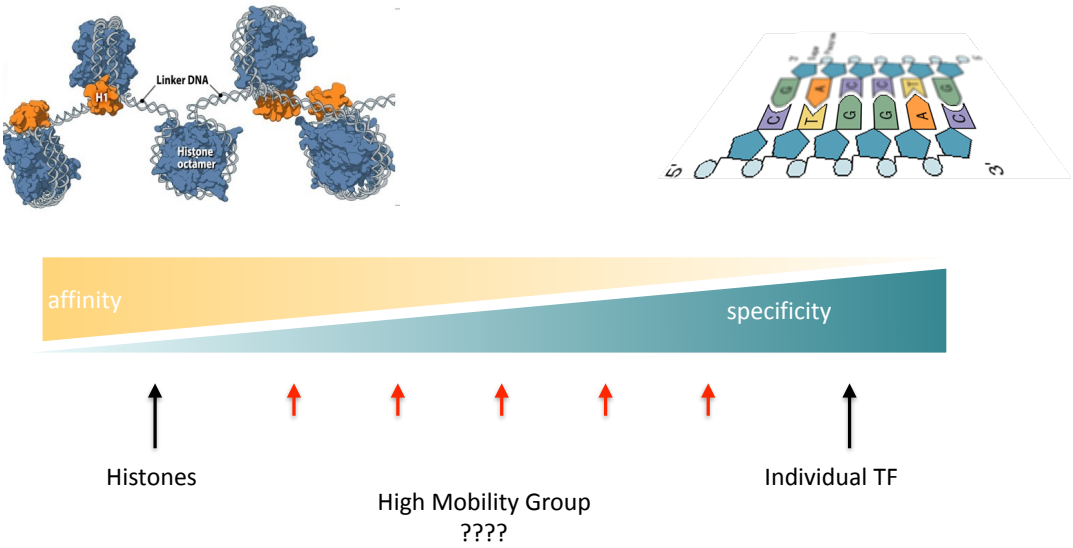


Figure 4-1 Histones have 150 bp protein-DNA surface and little sequence constraint. Transcription factors on the contrary recognize 6-20 bp sequence and at specific bases tolerate poorly eventual mismatches. What is the binding modality for HMG proteins is not known.

As a general consideration one has to remember that in vivo mammalian DNA is organized in nucleosomes, stretches of approximately 150 bp of DNA, wrapped around the two copies each of the four core histone proteins H3, H4, H2A and H2B.

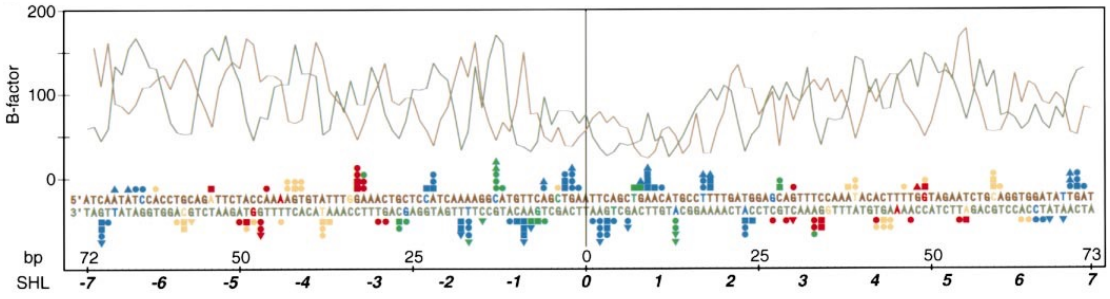
When we talk about chromatin we refer to nucleosomes, but also to nascent RNA and proteins that are bound either directly or indirectly to the genome, in the 8 µm wide cell nucleus. The estimated protein concentration in the nucleus is an exceedingly high: 100–400 mg/ml (Misteli, 2007). In this regime DNA-protein interaction strongly depend on avidity, and not only on the protein's absolute concentration and individual affinity. Indeed it has been shown that molecular crowding (for a review see Richter et al., 2007) is an important biological factor that facilitates stochastic interactions and makes them more efficient in vivo.

4.2.1 Histones and the formation of nucleosomes in vivo

There are approximately 30 millions of nucleosome per human cell (Bonaldi et al., 2002). The nuclear processes that take care of maintaining or renovating this complex structure (e.g. after each round of replication) are orchestrated by histone chaperones and remodelers (Burgess and Zhang, 2013).

Histones come as H3-H4 tetramers and are loaded by histone chaperones on DNA, giving rise to the tetrasome (Vlijm et al., 2015). Two H2A-B dimers are then added to form the nucleosome. Since the four core histone proteins are strongly basic, at physiological salt concentrations the histone octamer is stable only when wrapped around DNA. H1 protein, is the histone protein that binds to linker DNA, the 20 to 80 bp long region between nucleosomes (Woodcock et al., 2006). Its binding helps stabilizing compact arrays of nucleosomes shielding DNA backbone negative repulsion force.

In the seminal publication that first described the structure of the nucleosome, the contacts between the DNA phosphate groups and core histone proteins were highlighted (Luger, 1997). DNA phosphates have high mobility or are disordered when not contacted by histones (high B-factors), but effectively every 5 pb there is a histone DNA interaction (Figure 4-2).



(legend on next page)

Figure 4-2 DNA phosphate B-factors versus base pair. Symmetrical repetition of a 72 bp human α-satellite DNA across the dyad, wrapped around unmodified recombinant histones. The sequence of the DNA used is shown with corresponding B-factors ($Å^2$) plotted for the 59 phosphate group of each base. The contacts of the DNA phosphodiester chains with the histones are indicated: squares for main-chain hydrogen bonds; circles for side-chain hydrogen bonds, and triangles for hydrophobic bonds. Blue and green fro H3 and H4; yellow and red for H3-H4. The bases colored blue, green, red, and yellow indicate close proximity to an arginine side chain inserted into the minor groove. SHL, indicates the helical turns from the diad. Adapted from (Luger, 1997)

This piece of data highlights two properties of histones that are essential for their pervasive genome-wide distribution. First, the property of histone H3-4 to tetramerize before binding, assisted by histone chaperones is peculiar (Burgess and Zhang, 2013). This property allows initial contact with a large segment of DNA of around 50-60 bp. In sharp contrast, the majority of other DNA binders only form multimeric structures on the DNA scaffold (Jolma et al., 2015; Sainsbury et al., 2012).

The second property of histones is extensive non-sequence specific DNA contact: one nucleosome contains around 120 direct protein-DNA interactions and several hundred water-mediated ones (Davey et al., 2002).

On this basis it is possible to understand why nucleosomes are found throughout the genome. However certain genomic sequences are facilitated targets for their ability to maximize protein DNA contacts or disfavored due to their resistance to bending around the histone octamer. While in vitro nucleosome free energies can differ up to 1,000 fold (Thåström et al., 1999) in vivo not only the strongest theoretical binders are absent from the genome, but also occupancy can be modulated by the activity of chromatin remodelers (Segal and Widom, 2006).

Of note it is possible to find stretches of DNA of about 150 bp that contain only repeated A or T. This sequences, are known to be rigid and therefore devoid of nucleosomes in vivo (Raveh-Sadka et al., 2012).

It thus looks like that in vivo it is more important to avoid perfect histone octamer substrates than having nucleosome free DNA stretches. This is perhaps so because the replication and transcription machineries would collide and arrest at such strongly positioned nucleosomes.

Nucleosomes are thought to be important for genome biology for several reasons (C David Allis, 2014). First they wrap DNA in more compact volume units. They are also the repetitive sub-units of more compact genetic structures that are

refractory to transcription, like Polycomb repressive domains or HP1-positive heterochromatic compartments (discussed later). Additionally, the post translational modifications (PTMs) of histone tails are docking site for chromatin machineries and it has also been proposed that due to their left-handedness nucleosomes might also serve as reservoir of negative supercoiling (Naughton et al., 2013).

One well-documented role is modulation of TF binding (Barozzi et al., 2014; He et al., 2013) which is achieved thanks to free energies for nucleosomes 10 times higher than for single TF (Adams and Workman, 1995). In this regards we know for example that non-functional TF binding sites are embedded in regions of higher nucleosomal affinity (Field et al., 2011). On the other hand, nucleosome free regions tend to occur at clusters of TF binding sites (Valouev et al., 2011) and TFs that are able to probe new binding sites often recruit nucleosome remodeling activity (Ye et al., 2016).

4.2.1.1 Control of nucleosome stability and DNA-TF binding via histone modifications

Histone isoforms and histone PTM also play a role in modulating the competition between TF and histones for the DNA substrate and it looks like they do so mainly by stabilizing or destabilizing histone/DNA interaction (Henikoff, 2008). From work in yeast we know that acetylation of the globular domain of histones or histone variant H2A.Z are directly destabilizing nucleosomes (Tropberger et al., 2013; Watanabe et al., 2013). One opposite example is H3K36me3, which is deposited by transcription elongation and can be bound by a histone de-acetylase (Joshi and Struhl, 2005). This recruitment causes de-acetylation, thus preventing nucleosome destabilization and TF binding to cryptic promoters. Another example in the opposite direction is H3K9me3 which is able to recruit HP1, which in turns induces chromatin compaction (Hiragami-Hamada et al., 2016).

However when discussing the effect of chromatin compaction it is important to introduce the concepts of euchromatin, heterochromatin and in general of chromatin states (Bickmore and van Steensel, 2013).

Nuclear processes happening on the genome leave a mark of their action on chromatin by PTM of the molecules involved. One of the main process is transcription and generally high levels of histone acetylation and H3K4 methylation are detected in promoter regions of active genes (Bernstein et al., 2002; Roh et al., 2005). In addition to promoter regions, these modifications are also detected in intergenic regions and have been correlated with functional enhancers in various cell types (Heintzman et al., 2007). Methylation of H3K9 is involved in gene silencing (Bannister et al., 2001) and H3K27 methylation also correlates with gene repression (Boyer et al., 2006). In Figure 4-3a such bookmarking by chromatin modifications was exploited do divide the genome in distinct portions based on the functional state in a given cell type (Ernst et al., 2011). The number and composition of such chromatin types varies greatly depending on the level of clustering that one is aiming to achieve (van Steensel, 2011). One of the most conservative clustering based on histone modifications divides the genome into two functionally distinct states: euchromatin and heterochromatin. Interestingly, since the DNA sequence itself dictates most of the processes occurring on chromatin, these two types of chromatin are associated with a different sequence composition, and a different gene density (Martens et al., 2005). In heterochromatin for example the reduced number of genes per kilobase explains why marks associated with gene activation tend to be depleted. As a result nucleosomes are more compact and dense chromatin fibers are apparent by electron microscopy in interphase nuclei.

Of note, not only distinct chromatin types differ at the biochemical level, but physical differences are also observed. Whereas on the 1D genome, regions of active and inactive chromatin seem to alternate, in the 3D space of the nucleus they tend to coalesce: once a nuclear process starts, an associated sub-compartment also forms by stereo-specific interactions and it is energetically favorable to maintain it (Bancaud et al., 2009). This enhancement of molecular interactions by a self-governed biophysical process is generic and independent of specific biological functions and explains the existence some of some of the observed nuclear compartments (e.g. nucleolus, speckles, transcription factories, DNA damage foci, PcG bodies) (Figure 4-3b). Therefore when trying to analyze the impact that a specific histone modification or variant might have on

accessibility to DNA one has to keep in mind the additional levels of complexity at work. As discussed earlier, accessibility is determined by the compound action of nucleosome structure, chromatin associated proteins/modifications and effective protein concentration at the investigated nuclear compartment.

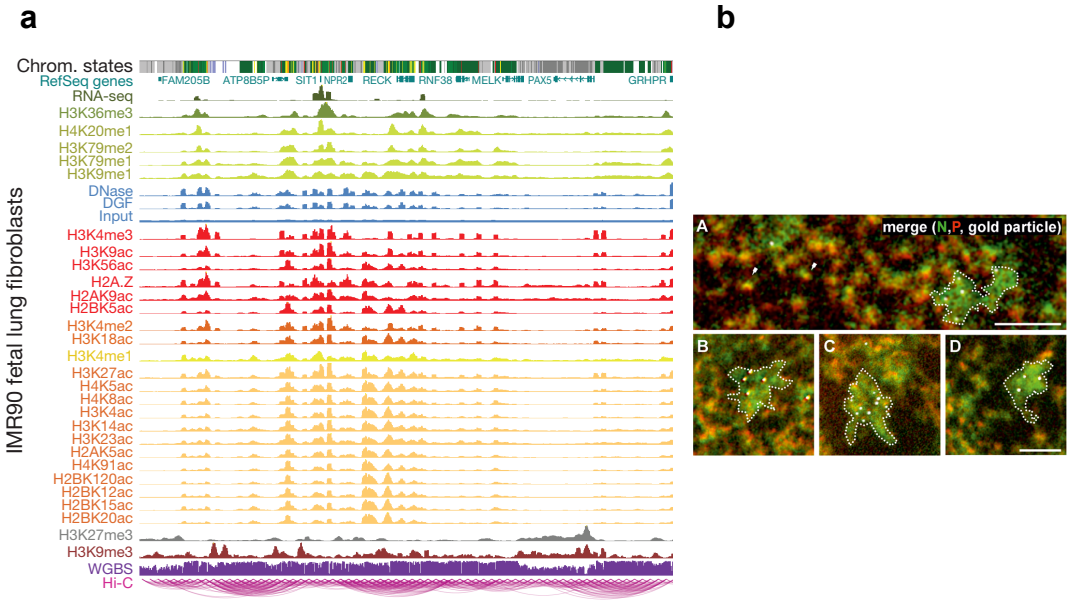


Figure 4-3 a) Different chromatin states can be called (colored top track) based on linear combinations of histone modification and other chromatin associated proteins or features. Image taken from (Meuleman et al., 2015); b) Images of transcription factories obtained using electron spectroscopic imaging. HeLa cells were permeabilized, nascent transcripts extended in BrUTP, and resulting BrRNA immuno-labeled with 5-nm gold particles; after sectioning (70 nm), images of endogenous phosphorus (red) and nitrogen (green), plus immuno- labeling gold particles (white), were collected and merged. (A) Five gold particles mark BrRNA in a nitrogen-rich factory (perimeter indicated by a dotted line). Absolute numbers of N and P atoms within this perimeter can be calculated using nearby nucleosomes as references (arrowheads). (B−D) Examples illustrating how poly- morphic factories are. Bars: 100 nm. From (Papantonis and Cook, 2013)

In general it is only recently that it became possible to examine the role of single histone PTM also because of the intrinsic difficulty in expressing in vivo histone mutants. Pioneering work in this direction points towards importance of histone marks for heterochromatin maintenance, and thus genome integrity, and cell identity rather than for transcriptional control (Jang et al., 2015; McKay et al., 2015). It could be that certain modifications are implicated in regulating TF access both at euchromatin and heterochromatin, but loss of the mark at regulatory regions appears to have less of a toxic effect.

4.2.2 Sequence specific DNA recognition: transcription factors

TF are defined as DNA sequence-specific binding proteins that do not have enzymatic activity or belong to the core transcriptional initiation complex. Most of them encode for factors that are either activator or repressor of transcription, either directly or through recruitment of cofactors (Biggin, 2011). The concerted action of all expressed transcription factors dictates the transcriptional state of a given cell. It is known that expression of some factor either alone or in simple combinations, is sufficient to drive trans-differentiation, reprogramming or stem cells differentiation (Takahashi and Yamanaka, 2016). Notable examples are Gata1, CEBP and MyoD and the reprogramming factors OSKM (octamer-binding protein 3/4, Sox2, Krüppel-like factor 4 and Myc). These factors are often referred to as master regulators as they can induce or revert cell fate decisions.

However the majority of the transcription factors seem not to share the same ability to autonomously recognize and bind their genomic targets.

If we could anticipate for every TF which targets are recognized in the genome, we would be much closer to correctly predict gene expression patterns. In order to complete a simplified transcription model, we would be left with determining the impact of binding on the targeted regulatory regions (activating or repressing) and then infer the induced enhancer-promoter connectivity maps (see section 4.1.1).

However already to predict which TF motif will be bound by a given TF is a daunting task (Kaplan et al., 2011). First of all, the availability of in vitro preference data for DNA sequences is not satisfactory yet, with hundreds of Zinc-finger TF specificities still waiting to be determined (Figure 4-4a).

The next layer of complexity lies in predicting the in vivo modulators of binding, introduced in the previous section, which can be divided in DNA-sequence related, TF related and mixed. Within the first category lies the positioning of nucleosomal arrays and the syntax (also referred to as lexicon or grammar) of a regulatory-region, which can determine cooperativity (Field et al., 2011). These two mechanisms in turns, autonomously determine DNA methylation and the torsional state of DNA (through transcription). TF related variables are on the other hand the levels of expression, presence of regulators and half-life. An

example of mixed modulation is the PTM state of a TF, which integrates both cues.

It is a complex mixture of all these levels of regulation that explains the homeostasis of nuclear functions and, when altered, the genetics of cellular defects (Kilpinen et al., 2013).

Syntax is probably the single most important determinant of TF binding.

TFBSs tend to cluster around CTCF or cohesin (Merkenschlager and Odom, 2013; Yan et al., 2013), two structural TFs that allow promoter/enhancer interaction. Cooperative binding leads to a nonlinear relationship between TF concentration and the degree of occupancy on specific enhancers. At the moment the prevailing idea posits that enhancers can function both as billboard continuum activators of transcription and also as cooperative digital switches (Slattery et al., 2014). Cooperativity is often associated with protein–protein interactions between TFs bound to adjacent sites, however there exist indirect types of cooperativity (recruitment of remodelers, displacement of a nucleosome, common co-factor). In order to test direct cooperativity in vitro there are now technologies, such as MITOMI and SELEX (Isakova et al., 2016; Jolma et al., 2015), that have the throughput and the sensitivity to give important insights into this area of investigation.

Another important mechanism of TF binding is presence or absence of a nucleosome occluding access to a cognate motif, and this mechanism is conserved from yeast to mammals (Field et al., 2011). Simple proof of the importance of this mechanism is that in vivo nucleosomes are depleted at regulatory regions but not at indivual TF binding sites, in light of the interplay with TF ensembles described above. Recent evidence support this notion by suggesting that nucleosomal array organization is both encoded in the DNA sequence and modulated by transcription factor binding (Barozzi et al., 2014).

Nucleosomes wrap DNA around the histone octamer and therefore only a region of approximately 6 bp (half of the B-DNA helical turn period of 10,5 bp) is available for initial DNA contact. It has been proposed that the unique properties of master regulators could lie in a different type of interaction with DNA (Soufi et al., 2015). There, authors show that master regulators are able to probe DNA sequences by recognizing partial motifs on only one side of the helix. Other TFs,

and in general those that recognize sequences of 10-12 bp (typically obligate dimers recognizing palindromic sequences), need to access both side of the helix for making a productive contact. Slightly against this hypothesis is the observation that TFs with a long motif (up to 21 bp) like CTCF or REST show a high fraction of bound sites, position nucleosomes and are not sensitive to chromatin changes (Stadler et al., 2011).

Following up on the issue of nucleosome displacement, another proposed mechanism of action for master regulators is recruitment of nucleosome remodelers (Voss and Hager, 2013). Collectively, these two evidences suggest a very important role for nucleosomes in modulating accessibility to DNA.

Another layer of control that is hard-wired in DNA sequence is DNA-methylation. The methylation state of a given fragment indeed is determined by the combined action of nucleosomal positioning (Baubec et al., 2015), CpG density and TF binding (Krebs et al., 2014; Stadler et al., 2011). A recent work from our laboratory has shown that this mechanism is responsible for modulating the binding of Nrf1 in mouse embryonic stem cells (Domcke et al., 2015). Additional evidence in different cell lines comes from modulation of CTCF binding at subsets of sites (Maurano et al., 2014).

Torsional state of the DNA is an emerging mechanism through which TF binding might be regulated. In turns torsional domains are determined by transcription and insulator proteins (Naughton et al., 2013). Its importance has been demonstrated in yeast by with the observation of nucleosomal eviction upstream of the replication fork and TF binding modulation (Gilbert and Allan, 2014; Langowski, 2015; Lia et al., 2003).

From the TF side, there are many examples that could illustrate the importance of TF concentration and activity. The best studied examples being nuclear receptors, SMAD, Nfkb and p53 however the advent of mass-spectrometry has opened new avenues in our capability of correct characterization and quantification of TFs at developmental times (Simicevic et al., 2013).

Also single cell tracking via fluorescent reporters allow microscopy based quantification of TF concentrations (Filipczyk et al., 2015; Hoppe et al., 2016).

The SymAtlas database contains the expression level by microarray of 873 TF, assessed for different human tissue types (Vaquerizas et al., 2009). In a give

tissue there are between 150 to 300 TFs expressed to levels detectable by the array, with 2 thirds of them being expressed to similar levels in all tissues and 1/3 generally expressed in a tissue specific fashion (Figure 4-4b).

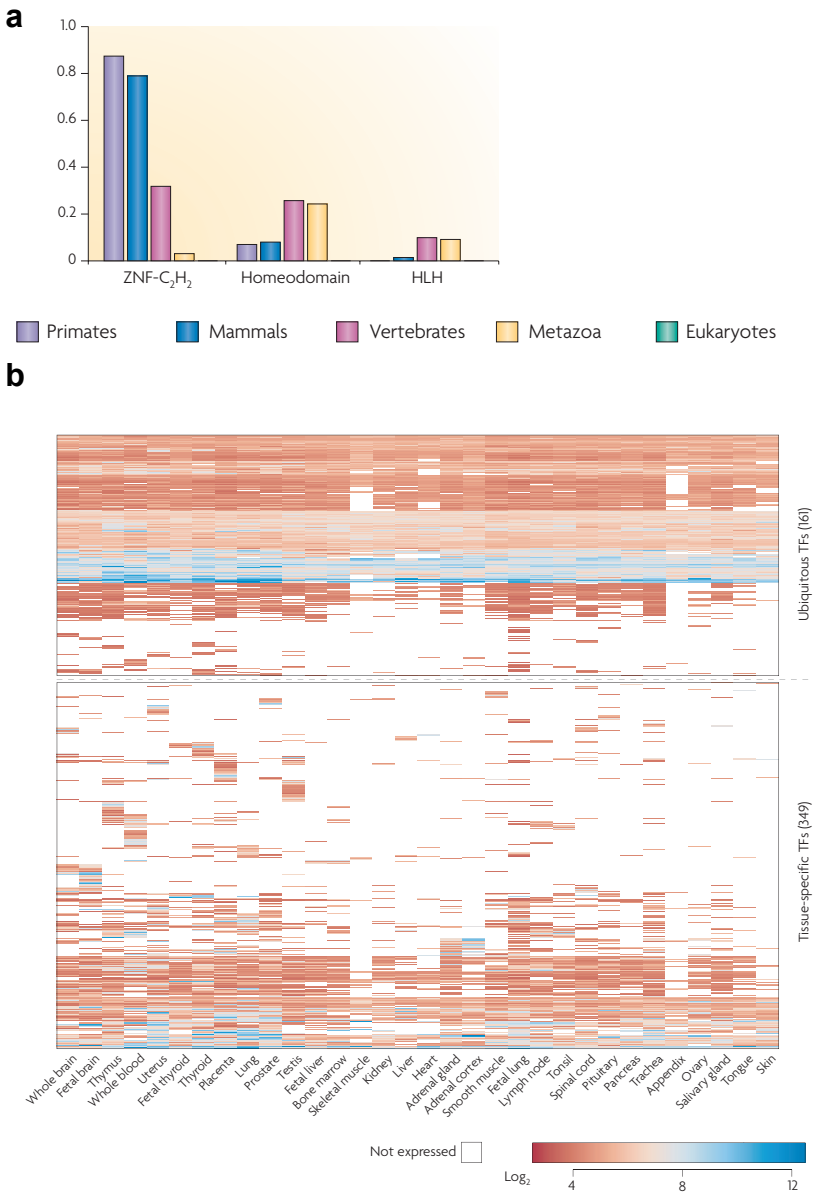


Figure 4-4 a) proportion of indicated class of TF in the different taxa; SymAtlas database showing expression levels of ~ 500 TFs in the different tissues; few transcription factors are active in only one cell type, with the majority being shared or with modulated expression. From (Vaquerizas et al., 2009)

As far as mixed (sequence and TF mediated) mechanism for control of TF binding, are for example the epigenetic state of the chromatinized targets, which

in turns integrates cues from the DNA sequence (number and type of TF bound at that particular cellular stage). It does so by modulating the level of chromatin compaction, the PTM of the TF and the methylation state of the DNA substrate (Tootle and Rebay, 2005).

What has been reported so far refers to TF in general. However TF-family specific properties are also known. For example HMG-box family of DNA binding protein bind preferentially to bended DNA structures, which have been implicated in indirect forms of cooperativity (Slattery et al., 2014).

In conclusion the forms through which TF can access their cognate sites are many and remain incompletely understood, however quantitative data is starting to accumulate.

Briefly a mention goes to the classes of DNA binding proteins that do not bind specific DNA sequence, like general transcription factors, UBF, MBD, CFP, etc. Investigation of the determinant of their binding is ongoing (Baubec et al., 2013; Lee, 2001) and, as far as HMG proteins are concerned, has been a main part of this thesis project.

4.3 Characteristics of mouse HMG proteins

HMG proteins were identified in 1973 by Ernest Johns, Clive Sanders and Graham Goodwin. With 0.35M NaCl extraction, they isolated a groups of proteins from calf thymus chromatin that rapidly migrated in polyacrylamide gel electrophoresis. Based on this they named those proteins "high-mobility group" proteins (Goodwin and Johns, 1973).

Fortuitously, later microscopy studies proved that such proteins are some of the most motile proteins in the nuclei of living cells (Harrer, 2004; Phair et al., 2004), which also showed that HMG proteins shared some biophysical properties.

From the biochemical perspective all have at least one DBD and most possess a negatively charged (acidic) carboxy terminal tail of varying length, which has been implicated in histone binding and intramolecular contacts (Sheflin et al., 1993).

According to updated inclusion requirements (Bustin et al., 1990) they are now defined as:

(1) extractable from chromatin using 0.35 M NaCl;
(2) soluble in 5% perchloric acid or tricloroacetic acid;
(3) < 30 kDa in molecular weight with a high content of charged amino acids;
(4) rapidly mobile in polyacrylamide gels;
(5) sensitive to extensive post-translational modifications such as phosphorylation, acetylation, and poly- ADP-ribosylation;
(6) tissue- and development-dependent expression.

According to this consensus in the mouse there are 2 Hmga genes, 4 Hmgb and 6 Hmgn genes (Table 4-1). They differ for the DNA binding domains, for their preferred substrates and for additional functions that are protein specific.

From a historical perspective, one of the first function attributed to HMGA1 was binding to major satellites (Strauss and Varshavsky, 1984). HMGB proteins (also known as HMG1-2 at that time) were identified as binders of Holliday junctions intermediates called cruciform DNA (Bianchi et al., 1989) and single stranded DNA (Bustin, 1999). HMGN are vertebrate specific and bind only nucleosmal DNA, and not free DNA or histones alone (Catez et al., 2002).

HMG motif protein	Functional motif	Root symbol	New name (canonical HMGs)	Old name (canonical HMGs)
HMG-box proteins	HMG-box	HMGB	HMGB1,2,..n	HMG-1,HMG-2
NBD proteins	NBD	HMGN	HMGN1,2,..n	HMG-14,HMG-17
ATH proteins	ATH	HMGA	HMGA1,2,..n	HMG-I/Y,HMG-C

Table 4-1 Nomenclature change after 2001. ATH, AT-hook domain; NBD, nucleosome binding domain. (Bustin, 2001). Source: www.informatics.jax.org/mgihome/nomen/hmg_family.shtml#G

Below is a resume of the major characteristics of HMGA-B proteins, that will be the focus of this work.

4.3.1 Expression

It has been calculated that in certain blood cells there are up to 1 million copies of HMGB1 (Čabart et al., 1995), which adds to 1 molecule per nucleosome and renders HMG proteins the second most abundant nuclear proteins (after histones).

As can be appreciated in Figure 4-5, cell-types in active replication tend to express HMG proteins higher than quiescent and terminally differentiated tissues, with the exception of HMGB4, which is confined to sperm tissue.

The high transcription of Hmg genes is backed up by fairly high stability of these short proteins. HMGB1 for example approaches the stability of histones, with a half-life of more than two cell generations (Begum et al., 1990).

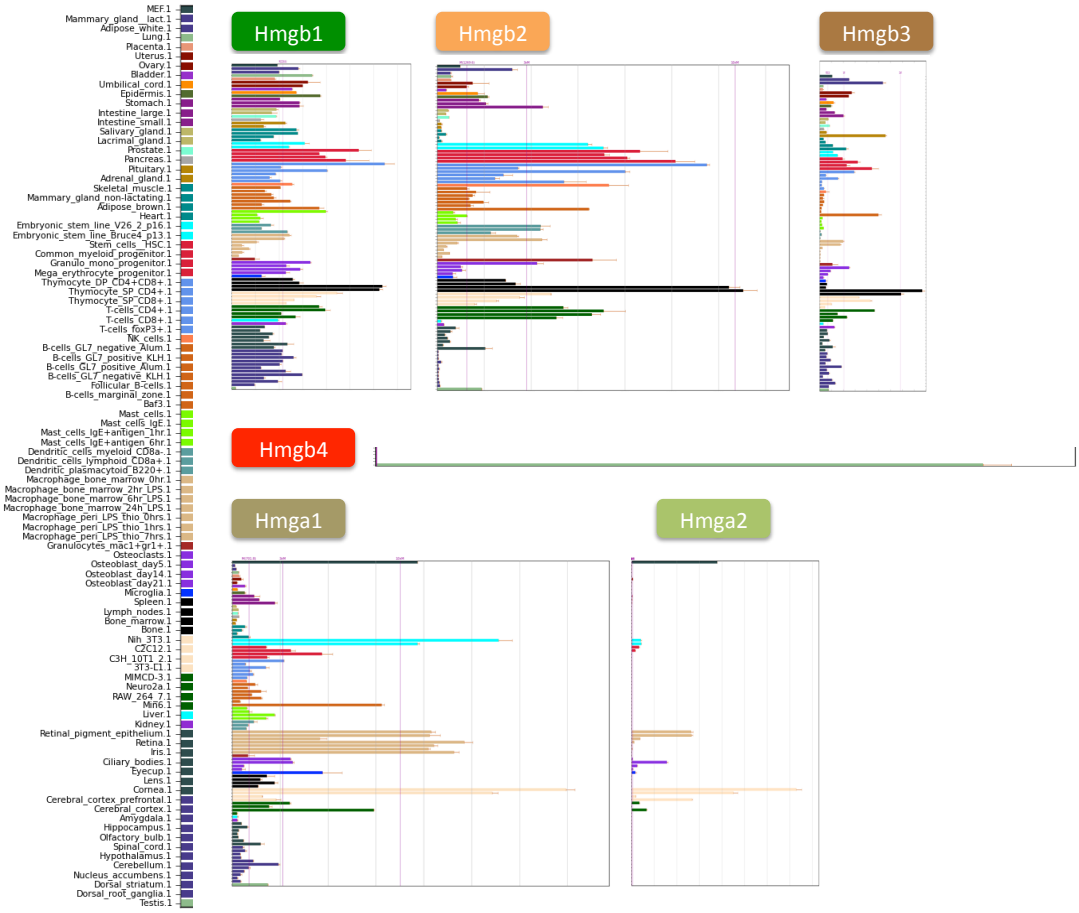


Figure 4-5 Expression levels for the mouse HMGA and HMGB proteins in different organs and developmental times. The different datasets were retrieved from microarray data from BioGPS (Wu et al., 2009) and were scaled to a common unit.

4.3.2 Amino acid sequence and structure of HMG proteins

HMGA1 and HMGA2 proteins possess 3 AT-hook DBD and an acidic tail (Figure 4-6). There is a high conservation between the DBDs. HMGA1 splice isoform b has a 11 aa reduced linker between AT-hook 1 and 2 similar to HMGA2. The AT-hook motif is a positively charged stretch of 9 amino acids containing the invariant repeat Arg-Gly-Arg-Pro. The AT-hooks domains explore the minor groove floor and make contact with the backbone of DNA (see Figure 8-2). HMGA1-2 proteins are unstructured proteins and they affinity to DNA increases additively when individual DBD are bound to the DNA target sequences (Frank et al., 1998). From bacteria to humans the AT-hook motif is conserved during evolution and is found in many, non-HMGA, proteins, the majority of which are transcription factors or are involved in chromatin remodeling (e.g. present in BRG1 and BRM, and CBX2) (Aravind and Landsman, 1998).

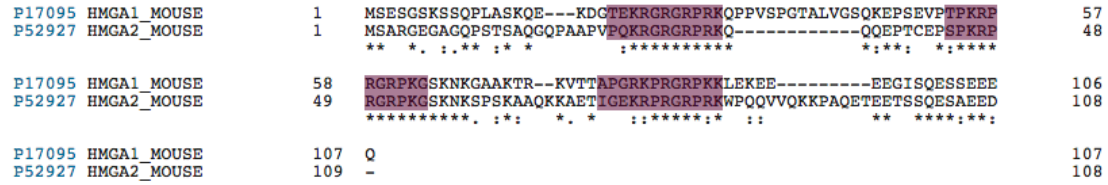
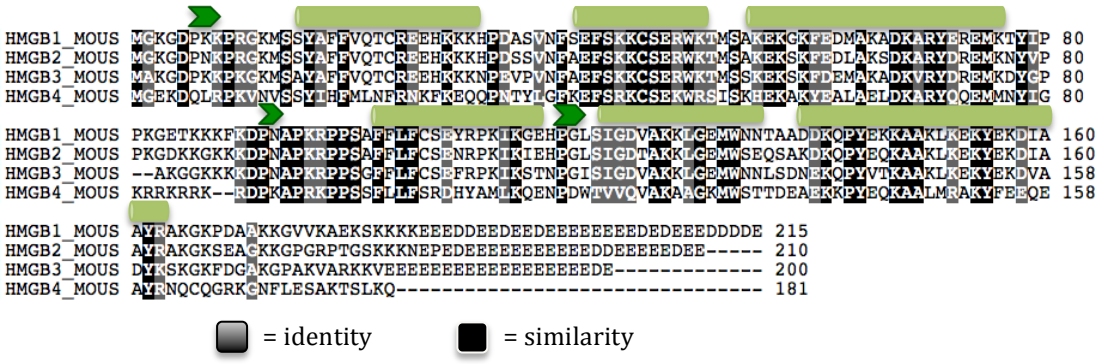


Figure 4-6 Alignment of the aa sequence of mouse HMGA1 and HMGA2. In purple are highlighted the DBDs. Conserved aa are marked with asterisks and similar aa with "." or ":" symbols. The linker between the first two AT-hooks is absent in HMGA1 isoform b.

HMGB proteins all share two HMG-box DBDs of approximately 70 aa (Figure 4-7). Such domains are constituted of 3 alpha helices and short beta-sheet or linkers. The bulky aromatic side chains of AA phenylalanine, tyrosine and isoleucine at the beginning of the alpha helices are important for the in vitro DNA bending properties of these proteins (Sanchez-Giraldo et al., 2015)(Thomas and Travers, 2001). HMGB4 lacks the long acidic tail characteristic of the other HMGB members. The NLSs of HMGB1-2-3 are two and are located between the first two alpha-helices (bipartite) and just before the acidic tail (only for HMGB1-2). HMGB1 and HMGB2 proteins shuttle in and out of the nucleus to the cytoplasm by simple diffusion and active transport, with a tendency to nuclear accumulation (Pallier et al., 2003).

HMGB1 is highly conserved in mammalian evolution with amino acid sequences virtually identical (>99%), implying similar biological functions.



(legend on next page)

Figure 4-7 Alignment of the aa sequence of mouse HMGB proteins. In green beta-sheet and alpha helices for the two HMG-box of HMGB1 are indicated. Note the lack of amino acidic tail for HMGB4 and the different NLS at position 80.

4.3.3 Evidence of association with DNA and chromatin

Initial evidence for association with chromatin dates back to the discovery of the proteins themselves, as they are defined by being bound to chromatin and being extracted at 0.35 NaCl (See previous sections). However numerous studies have tried to identify the in vivo substrates of HMG proteins with mixed fortune.

4.3.3.1 In vitro and in vivo evidence for HMGA proteins association with DNA

In vitro, the HMGA1 protein binds preferentially to the minor groove of short stretches of A/T-rich B-form DNA via recognition of structure rather than nucleotide sequence (Reeves and Nissen, 1990).

Full-length proteins and DNA-binding domain(s) alone also bind to synthetic four-way junction structures, to non-B-form structures in supercoiled plasmids, and to distorted regions of DNA found on isolated nucleosome core particles (Hill et al., 1999; Reeves and Wolffe, 1996).

In vivo, they localize to DAPI dense foci. Of note the true reason for brighter staining of DAPI dense foci is not yet known. Whether it is caused by an increased DNA compaction of DNA or by stronger DNA staining due to the dye affinity, or combinations of the two. Indeed DAPI, Hoechst and other DNA dyes display in vitro high affinity for AT-rich DNA, like AT-hook domains, sharing with them similar mechanisms of substrate recognition (Reeves and Nissen, 1990; Wilson et al., 1990).

Evidence accumulated over the years, implicates HMGA1 in the formation and stabilization of activating protein complexes at the enhancer of the IFN-β and IL2-Rα involving the interaction of HMGA1 with other transcription factors (Merika and Thanos, 2001).

Looking at a more global role for HMGA proteins, it has been demonstrated that treatment of cells with minor groove binding chemicals prevents complete condensation of metaphase chromosomes (Radic et al., 1992). Also, HMGA proteins are in vivo substrates for CDC2 kinase and colocalize with histone H1 and TopoisomeraseII at scaffold attachment regions (A/T-rich DNA sequences that constitute the structural backbone of metaphase chromosomes) (Zhao et al.,

1993). Overexpression of HMGA proteins induces apoptosis in normal cells (Fedele et al., 2001), however in cell lines that have passed the senescence barrier this has anti-apoptotic affects. The observation that HMGA1 localizes at senescence-associated heterochromatic foci in fibroblast (Zhang et al., 2007), which contain hypo-acetylated histones and where H1 is excluded might be related.

More recently a ChIP-sequencing study located HMGA2 in the colon cancer cell line HCT116 (Figure 4-8). 49 DNA fragments were analyzed for their AT-content and showed a significantly higher AT-content than the average of the human genome (Winter et al., 2011).

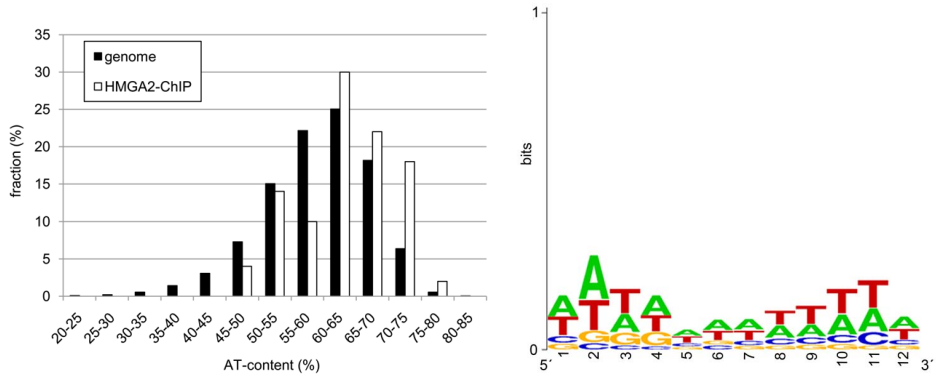


Figure 4-8 Left, distribution of sequence content of the 49 clones analyzed in this study after HMGA2 ChIP ; Right, in vivo determined consensus from, the information content is fairly low and the nucleotide composition degenerate (Winter et al., 2011)

4.3.3.2 In vitro and in vivo evidence for HMGB proteins association with DNA

The first DBD of HMGB1, also called box-A has been crystallized in complex with distorted DNA templates (Lippard et al., 1999) and is shown in Figure 4-9. The box-B has been shown to have even higher affinity for damaged DNA and induce an angle of up to 95° on the DNA (Thomas and Travers, 2001). Specific amino acids have been identified as being responsible for the strong affinity for distorted DNA substrates Figure 4-9. HMGB does not have reported sequence specificity and thus plausible in vivo ligands that spontaneously adopt this structures are UV-induced pyrimidine dimers, to which HMGB1 binds efficiently in vitro (Pasheva et al., 1998). However strong bends on the DNA helix are also

caused by canonical HMG-box TF like Sox2 (Scaffidi and Bianchi, 2001), therefore it cannot be excluded that HMGB protein recognize bended DNA, for example at the nucleosome entry-exit point.

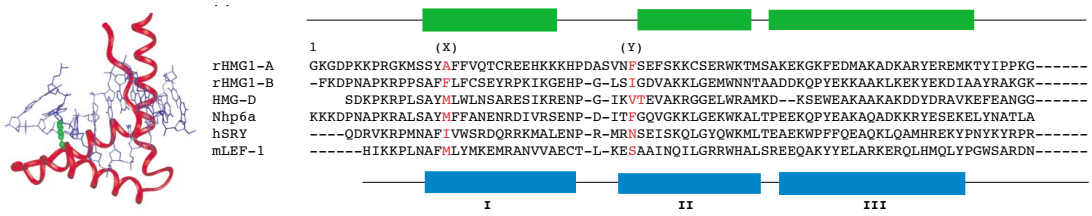


Figure 4-9 Left, The structure of the box-A of HMGB1 in complex with Cis-platinated DNA. Sharp angle of 60° is imparted to DNA. Right, Conserved amino acid are marked in red that are thought to be important for the intercalation and DNA bending in vitro. First 2 lines are rat/mouse sequence for the two boxes. Alignment with Drosophila and Yeast homologs HMG-D and Nhp6a. For comparison also sequence specific hSRY and mLEF-1 are depicted. From (Thomas and Travers, 2001)

For this reason, a model has started to prevail that implicates HMGB proteins in the loading of nucleosomes and modulation of their stability (Joshi et al., 2012; Stros, 2010; Travers, 2003; Watson et al., 2014).

Additionally a role for stabilization of enhancer complexes has emerged after studies conducted on selected loci and on reporter plasmids (Das et al., 2004; Roemer et al., 2008; Verrijdt et al., 2002). In support of this hypothesis a recent publication describes by ChIP-seq HMGB2 at the promoters of two cancer derived cell lines (Redmond et al., 2014).

A third mechanism implicates HMGB in the control of protein-DNA interaction, the so-called “hit-and-run” mode of action. According to this model HMGB proteins facilitate stable binding of other TFs to regulatory regions but immediately dissociates from the ternary complex, which remains firmly bound (Agresti and Bianchi, 2003). Not all interactions of HMGB proteins are of the hit-and- run variety: in the case of the BHLF-gene, HMGB facilitates the formation of an enhanceosome (Ellwood et al., 2000). However the clearest example of stable HMGB1 complex formation in vivo occurs during apoptosis when the dynamic movement of this protein is completely arrested (Scaffidi et al., 2002). The signal for such binding is unknown but it has been suggested that it might be due to either HMGB recognition of (and binding to) the hypoacetylated N-terminal histone tails. The apparent biological function of this binding in apoptotic cells is

to prevent the protein from leaking out of dying cells and triggering an inflammatory response (Bianchi and Manfredi, 2004).

4.3.4 Post translational modifications

The stability of HMG proteins, their localization and interaction with other proteins and DNA, is highly regulated by post-translational modifications (PTM), including ADP-ribosylation, acetylation, methylation and phosphorylation.

This is reminiscent of histone PTM and there is indeed evidence of some overlap among the enzymes that modify the two groups of proteins.

HMGAs are phosphorylated more than any other protein in the nucleus and phosphorylation decreases DNA binding affinity (Lund et al., 1985).They are also acetylated at several residues and methylated at Arg in the DBDs in a manner that affects DNA binding (Cleynen and Van de Ven, 2008).

For HMGB1, phosphorylation, acetylation and ADP-ribosylation have been implicated in cytosolic export and excretion (Zhang and Wang, 2008).

4.3.5 Pseudogenes

Important genes for cell survival are often found in multiple copies, for example ribosomal genes and histone genes. HMG proteins, apart from having multiple paralogs are also present in multiple copies in the genome as pseudogenes.

One of the main sources of pseudogenes is gene retrotransposed copy (RTC) mechanism. It is known that short transcripts of abundant proteins are often subject to retro transposition (Gonçalves et al., 2000). This process produces RTC, which are cDNAs embedded in some repeat elements, have a DNA encoded polyA and are normally devoid of a promoter. Due to the fact that these RTC are often not expressed many mutations accumulate thus compromising the possibility of generating functional proteins.

In humans, an extensive study has been carried out on HMG proteins and 219 copies could be identified for HMGA-B-N, with sequence similarity ranging from 64% to 98% (Strichman-Almashanu et al., 2003). These studies also showed that approximately up to 10% of HMGB1 transcripts align to RTC (57 in total) and that 3 spliced isoforms of HMGA1 aligned to the HMGA1 RTC (6 in total).

Of note, it has been recently suggested that in human overexpression of pseudogenes might impair microRNA mediated control of Hmga1 gene expression (Esposito et al., 2015).

In the mouse at least 28 sequences show high homology to HMGB1 cDNA however only half of them can be mapped uniquely to the mm9 genome (Table 4-2).

Symbol	Description	Chromosome
Hmgb1	high mobility group box 1	5
Hmgb1-ps1	high mobility group box 1, pseudogene 1	11
Hmgb1-ps10	high mobility group box 1, pseudogene 10	13
Hmgb1-ps11	high mobility group box 1, pseudogene 11	13
Hmgb1-ps2	high mobility group box 1, pseudogene 2	X
Hmgb1-ps4	high mobility group box 1, pseudogene 4	19
Hmgb1-ps5	high mobility group box 1, pseudogene 5	3
Hmgb1-ps6	high mobility group box 1, pseudogene 6	16
Hmgb1-ps7	high mobility group box 1, pseudogene 7	7
Hmgb1-ps8	high mobility group box 1, pseudogene 8	10
Hmgb1-ps9	high mobility group box 1, pseudogene 9	13
Hmgb1-rs10	high mobility group box 1, related sequence 10	8
Hmgb1-rs11	high mobility group box 1, related sequence 11	17
Hmgb1-rs12	high mobility group box 1, related sequence 12	18
Hmgb1-rs13	high mobility group box 1, related sequence 13	X
Hmgb1-rs14	high mobility group box 1, related sequence 14	X
Hmgb1-rs15	high mobility group box 1, related sequence 15	11
Hmgb1-rs16	high mobility group box 1, related sequence 16	9
Hmgb1-rs17	high mobility group box 1, related sequence 17	8
Hmgb1-rs18	high mobility group box 1, related sequence 18	4
Hmgb1-rs19	high mobility group box 1, related sequence 19	
Hmgb1-rs20	high mobility group box 1, related sequence 20	
Hmgb1-rs21	high mobility group box 1, related sequence 21	
Hmgb1-rs22	high mobility group box 1, related sequence 22	
Hmgb1-rs23	high mobility group box 1, related sequence 23	
Hmgb1-rs5	high mobility group box 1, related sequence 5	13
Hmgb1-rs6	high mobility group box 1, related sequence 6	17
Hmgb1-rs8	high mobility group box 1, related sequence 8	6
Hmgb1-rs9	high mobility group box 1, related sequence 9	8

Table 4-2 List of the HMGB1 pseudogene sequences found in the mouse genome. From (Strichman-Almashanu et al., 2003)

For the HMGA1 protein, at least 2 pseudogenes can be aligned to the mouse genome, and three splice isoforms are known for the Hmga1 gene itself (Cleynen and Van de Ven, 2008; Strichman-Almashanu et al., 2003).

4.3.6 Phenotypes associated with genetic deletion and overexpression

Below is a list of the described phenotypes associated with various HMG protein alterations. For the vast majority of the observed phenotype a link between genotype and phenotype has not been found yet.

Organism	HMG	HMG level	Phenotype
Mouse	Hmga1	Heterozygous	Impaired spermatogenesis; cardiac hypertrophy
Mouse	Hmga1	KO	Type 2 diabetes; cardiac hypertrophy
Mouse	Hmga1	KO	Impaired lymphohematopoietic differentiation of ESC
Mouse	Hmga2	KO	Pygmy; reduced fat tissue; impaired spermatogenesis
Mouse	Hmga2	Overexpression	Effects on myogenesis in ESC
Mouse	Hmga1-2	DKO	Pygmy(25%); reduced growth rate
Mouse	Truncated Hmga2	Transgene overexpression	Obesity
Human	HMGA1	Reduced expression	Type 2 diabetes
Mouse	Hmgb1	KO	Animals die within 24h because of hypoglycemia
Mouse	Hmgb2	KO	Defect in spermatogenesis
Mouse	Hmgb3	KO	Erythrocythemia
Human	Hmgb2	KD	Erythrocytopenia
Human	Hmgb3	KD	Erythrocythemia
Xenopus	Hmgb3	KO	Reduction in eye and brain size
Xenopus	Hmgb3	Overexpression	Increased eye and brain size

Table 4-3 Phenotypes of the indicated HMG alteration. Adapted from (Hock et al., 2007)

Recently a high throughput KD study identified HMGB2 and HMGA1 as important for correct genome compartimentalization (Shachar et al., 2015).

HMGB proteins are not essential for cell viability in vitro but cell tend to suffer upon DNA damage from higher induced toxicity, which led to the hypothesis that HMGB1 might also have a direct role in DNA repair (Yumoto et al., 1998). Reports on Hmgb1 KD reducing the histone content and affecting nucleosomal structure have also appeared (Celona et al., 2011).

For HMGA proteins increasing evidence indicates that deregulation and rearrangements of HMGA proteins are a hallmark of both of malignant and benign neoplasia. Table 4-4 summarizes the major associations discovered so far implicating HMGA proteins as oncogenic agents. At the same time HMGA overexpression sensitizes cancerous cells to killing by various genotoxic agents such as UV light, cisplatin, hydrogen peroxide, menadione and methy-methanesulfonate (Reeves, 2010).

Human disease	HMGA1 involvement
Bladder cancer	Overexpression
Breast cancer	Overexpression
Colorectal cancer	Overexpression; positively regulates Wnt/b-catenin signaling
Head and neck cancer	Overexpression
Leukemia	Overexpression; Cmyc target
Kidney cancer	Overexpression
Liver cancer	Overexpression
Lung cancer	Overexpression; promotes transformation
Glioblastoma/Neuroblastoma	Overexpression
Pancreatic cancer	Overexpression; promotes cellular invasiveness and metastatic potential
Prostate cancer	Overexpression; involved in chromosomal re-arrangements
Gastric cancer	Overexpression; let7-downregulation
Thyroid cancer	Overexpression; regulates expression of miR-603 and miR-10b
Cervix cancer	Overexpression
HIV infection	Co-factor for integration, transcription and splicing
Human papovavirus JC infection	Co-factor for transcription
Epstein Barr virus infection	Co-factor for transcription
Herpes Simplex virus 1 infection	Co-factor for transcription
Alzheimer's disease	Involved in presenilin-2 pre-mRNA exon-skipping

Table 4-4 Involvement of human HMGA1 in various disorders. In general overexpression is associated with benign and malignant tumors. From (Benecke et al., 2015)

5 Aim of the work

Aim of the thesis was to gain insights into the binding preferences and function of the High mobility group class of proteins.

This class comprises proteins that are very abundant in dividing cells, yet poorly understood with respect to nuclear function and binding modalities in vivo.

Histones tend to bind DNA in manner that reflects the biophysical properties of a DNA stretch (flexibility) whereas transcription factors (TF) in a sequence dependent manner. At the time we begun our study, no genome-wide data was available for HMGB and HMGA proteins that could inform on the nature of their binding to DNA. Learning the genomic location of HMG proteins would help us revise the many functional models that have been proposed, which are mainly based on in vitro data and sporadic in vivo observations.

HMGN proteins were not included in our study because genomic maps were already available for mice and humans (Cuddapah et al., 2011; Deng et al., 2013; Zhang et al., 2016) and their nuclear function has been reviewed elsewhere (Kugler et al., 2012). A brief discussion of those results will be made in the result section while commenting on HMGB data.

As a pilot study, we also analyzed the location of additional TFs belonging to different TF classes. This data served mainly as a control to confirm previously described TF binding patterns and, by doing so, to evaluate the technical feasibility of the study.

6 Materials and methods

The RAMBiO approach has been already described (Baubec et al., 2013). Here below is a summary of the relevant procedures adopted in this work.

6.1 Cloning and generation of cell lines harboring biotin tagged TF

Where possible cDNAs were amplified from a random hexameres reverse transcription cDNA library (Superscript III, Invitrogen) generated from RNeasy extracted total RNA (QIAGEN, 74104). cDNAs were then cloned into pL1-CAG-bio-MCS-polyA-1L. The two inverted L1 Lox sites allowed CRE mediated integration into a unique genomic site. Recombination allows excision of the negative selection marker TK (Figure 6-1). Gancyclovir (6 µM) resistant clones are selected and tested for direction of the integration through junction-PCR. The parental cell-line contains and homogenously expresses BirA-V5 biotin ligase, which leads to stable biotinylation of the bioTF (expressed under the CAG-promoter) throughout differentiation (Baubec et al., 2013). Importantly, in all cases of HMG expression there was no phenotypic and growth alteration, neither in ESC nor at NPC stage (Figure 6-1b).

Cells were cultivated on feeder cells or 0.2% gelatine coated dishes. ES cell growth medium consisted of DMEM (Invitrogen) supplemented with 15% fetal calf serum (Invitrogen), 13 nonessential amino acids (Invitrogen), 1 mM L-glutamine, LIF, and 0.001% beta-mercaptoethanol. Differentiation was performed as previously described (Bibel et al., 2004).

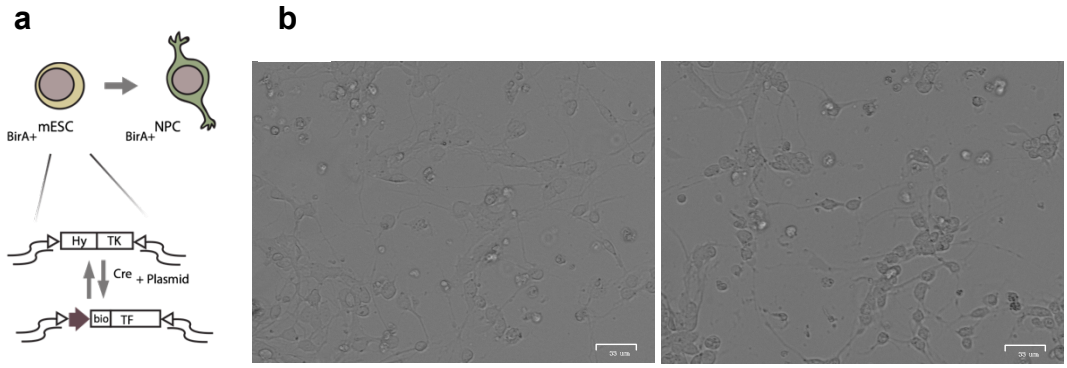


Figure 6-1 a) RAMBiO protocol for the generation of biotinylated TF of interest; b) Plated neuronal progenitor (day 8) cells for the Parental cell line (Left) and a HMGA1 expressing clone (Right). No differences can be appreciated at microscopy inspection.

Protein expression was initially screened by western-blotting (data not shown) with streptavidin (SAV) coupled horseradish peroxidase (HRP). Nuclear proteins from ESC were enriched by nuclear fractionation and IP. Streptavidin-biotin pull-downs were performed with preblocked (0.1% cold fish skin gelatine) 30 ml Streptavidin-M280 magnetic beads (Invitrogen) in HENG buffer, 150 mM NaCl, at 4°C overnight. Streptavidin magnetic beads were washed three times each 10 min with HENG buffer, 250 mM NaCl, 0.3% NP40, 1 mM DTT, and protease inhibitors at 4°C. IPs and 5% inputs were resuspended in Laemmli buffer prior to SDS-PAGE and western blotting (WB) to PVDF membranes.

In the images presented in this thesis protein specific antibodies were used for blotting on whole cell extract (TNN extraction buffer) WB. Membranes were blocked with 5% milk or 5% BSA for detection with antibodies or Streptavidin-HRP, respectively.

6.2 Streptavidin-fluorescence and Immuno-fluorescence Microscopy

PBS Cells suspensions were place on poly-L-lysine for 10 minutes, fixed for 10 min in 3% PFA and permeabilized in 0.1% NaCitrate and 0.1% Triton X-100. After 30 min blocking with 0.1% Tween20, 3% BSA (w/v) and 10% normal goat serum in PBS, detection was performed with Streptavidin-AF568 (ThermoFisher) or primary antibodies over night at 4_C. Coverslips were washed with 0.1% Tween20 and 0.25% BSA (w/v) in PBS (cells stained with primary antibody where incubated with secondary antibody at room temperature for 30 min). DAPI counter staining was performed for 10 min at room temperature. Images were taken using a Zeiss Z1 epifluorescence microscope. Image analysis was done with ZEN (Zeiss) and final images were assembled in Illustrator (Adobe).

Localization of HMGB1 was contrasted to HMGB1 in HeLa with Abcam ab206896, HMGB2 in PC-12 cells with Abcam ab124670 and HMGB3 in skin fibroblasts with R&Dsystems MAB55071 (online datasheets).

6.3 CRISPR design and KO strategy

Desing tools used: http://crispr.mit.edu and http://www.e-crisp.org/E-CRISP/

A pX330 plasmid expressing CRISPR-Cas9 and guide RNA together with a reporter expressing Puromycin-2A-mCherry were co-transfected in ESC. On the

following day, Puromycin (2 µg/mL) was added and cells were kept under selective media over night. Media was refreshed the next day and after single cell plating, clones were isolated. After PCR amplification of a 700 bp region centered on the CRISPR guide indels were analyzed with TIDE (Brinkman et al., 2014) and confirmed by Sanger sequencing (Figure 6-2). WBs with specific antibody were performed to confirm absence of targeted proteins. KO strategy relied on introducing frame-shift mutations in the coding sequence of Hmgb1 and Hmga1. We targeted intron-exon junctions in order to avoid off-targets caused by the presence of pseudogenes.

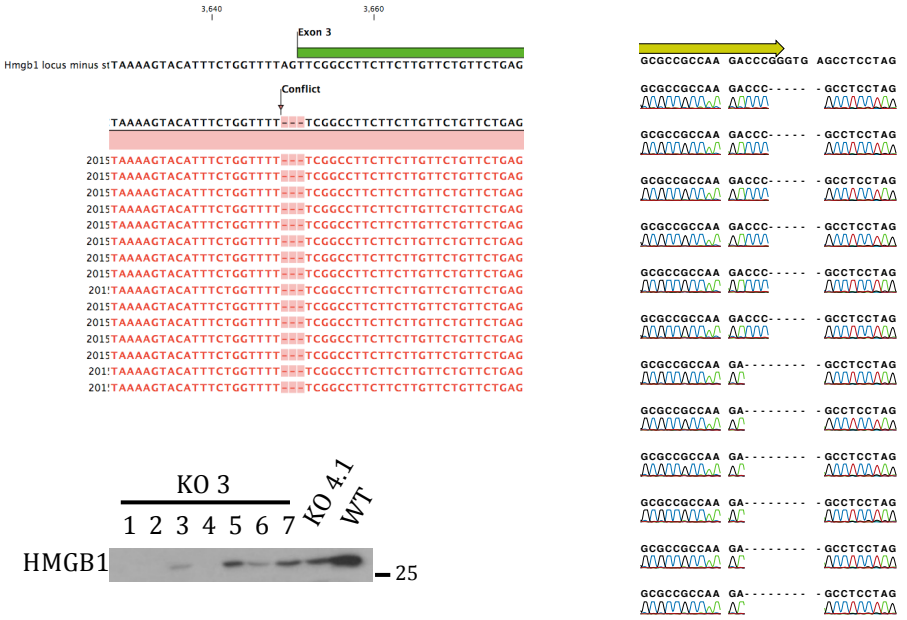


Figure 6-2 Left, Hmgb1 locus and Hmga1 locus (right) sanger sequencing of PCR products demonstrating indels causing frameshift of the indicated open reading frames. Bottom. WB on equal amounts of cellular lysates from clones of two different KO experiment (KO 3 / 4). Clone 3.1 was used for characterization and add-back experiments. Blotting with HMGB1 antibody. See 7.3.6 for HMGA1 WB.

6.4 bioChIP and Sequencing

Enrichment of bound-DNA is achieved via streptavidin (SAV) pull-downs (bioChIP), in a very similar way to Chromatin Immuno Precipitation (ChIP) experiments, however here more stringent washes are possible due to the nM interaction between biotin and SAV. Libraries for next-generation sequencing are generated with extracted DNA from the IP and input (50 µl) fraction.

Additionally, we made two other important controls, namely biotin tagged monomeric GFP pull-down (freely diffusing in the nucleus) and pull-down in the

parental cell line devoid of the expression construct. For both these experiments traces amount of DNA below lower detection limit of Qubit dsDNA HS Assay Kit (ThermoFisher) could be retrieved after pull-down. Library preparation for ChIP-seq failed for the empty cell line cell line, highlighting that our protocol does not suffer from contaminations. Library preparation for the GFP sample was possible and gave us the background random contact map of a ~ 30 kDa protein with the genome (as captured by formaldehyde fixation).

For cross-linking and chromatin extraction, cells were fixed for 10 min with 1% formaldehyde at room temperature and incubated for 10 min on ice in presence of 125 mM glycine. Cells were harvested and treated for 10 min with 10 mM EDTA, 10 mM TRIS, 0.5 mM EGTA, and 0.25% Triton X-100 and 10 min in 1 mM EDTA, 10 mM TRIS, 0.5 mM EGTA, and 200 mM NaCl with subsequent lysis in 50 mM HEPES, 1 mM EDTA, 1% Triton X-100, 0.1% deoxycholate, 0.1% SDS, and 150 mM NaCl for 2 hr on ice. Crosslinked chromatin was subjected to sonication in a Bioruptor instrument (Diagenode). Streptavdin-M280 magnetic beads were blocked for 1 hr with 0.1% cold fish skin gelatin and 100 ng tRNA. 150–250 mg chromatin were precleared with protein-A Dynabeads (ThermoFisher, #10001D) and incubated with 40 µl blocked Streptavidin-M280 (ThermoFisher, #11205D) magnetic beads overnight at 4°C. Beads were washed with two rounds of 2% TE-SDS, once with 500 mM NaCl sonication buffer, once with 250 mM LiCl, 1 mM EDTA, 0.5% NP-40, 0.5 Deoxycholate, 10 mM TRIS, and two rounds of TE. Beads were then transferred to a fresh tube and bound chromatin was eluted in elution buffer containing 1% SDS and 100 mM $NaHCO_3$. Beads and eluate were treated with RNaseA for 30 min at 37°C and proteinase K for 3 hr at 55°C and were then decrosslinked overnight at 55°C. DNA was purified with QIAquick PCR Purification Kit (QIAGEN, #28104).

6.4.1 Library preparation protocols

Libraries from extracted DNA were prepared according to the manufacturer's protocol for NEBNext ChIP-Seq Library Prep Master Mix Set for Illumina (New England BioLabs, #E6240). Input DNA was measured using NanoDrop 3300 Fluorospectrometer (Witec AG) and Qubit dsDNA HS Assay Kit (ThermoFisher). Samples were end-repaired, dA-tailed and adapter ligated Size-

selection was performed using Agencourt AMPure XP beads (Beckman Coulter, # A63880) before PCR amplification with NEBNext Multiplex Oligos for Illumina (New England BioLabs, #E7335). PCR amplification was performed for 6 to 12 cycles using indexed primer and cycling conditions according to Illumina recommendations. Adapter-ligated and amplified DNA was purified using AMPure XP beads. Before pooling size distribution was checked on Agilent Bioanalyzer 2100 using Agilent High Sensitivity DNA kit (Agilent technologies, #5067-4626).

Alternatively library preparation was performed according to NEBNext Ultra DNA Library Preparation Kit (New England Biolabs, #E7370L).

For Foxo1/3, Smad3/4, Sox2 samples library preparation was performed with gel size-selection (250-300 bp).

For all samples sequencing was performed on an Illumina HiSeq 2500 machine (50 bp read length, single-end, according to Illumina standards). Normally, four libraries were pooled at equimolar ratio in one sequencing lane, yielding 40M unique reads per sample.

6.4.2 Variation in bioChIP protocol

6.4.2.1 Low temperature extended crosslinking

A recent paper described HMGB2 genome-wide binding profiles in human breast cancer cell lines (Redmond et al., 2014). In that study, by adopting a 1h crosslinking protocol at 4°C, the investigators observed by antibody ChIP-seq similar HMGB2 enrichments at active regulatory regions but with a higher dynamic range. We replicated that crosslinking protocol with the bioChIP approach. After library-size normalization we could not see any improvement in signal-to-noise ratio Figure 6-3. Therefore we conclude that their observation is likely antibody dependent or cannot be reproduce with the more stringent bioChIP protocol.

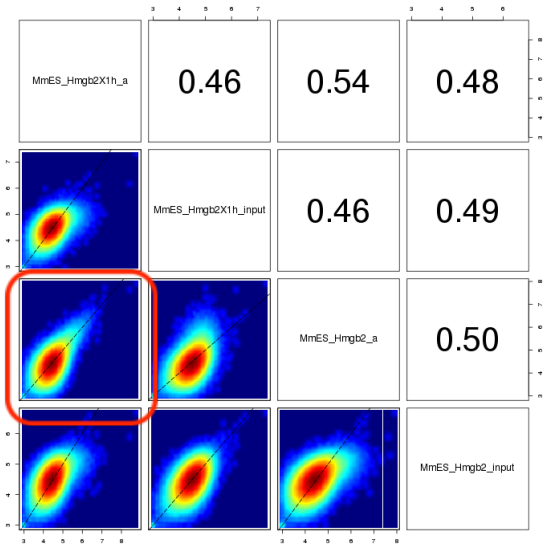


Figure 6-3 Scatterplots and Pearson's correlation coefficients for IP and input HMGB2 samples processed either with standard bioChIP protocol (lower two) or 1h 4°C (upper two). Shown are Log2 library size normalized read counts over 500 bp windows on Chr1.

6.4.2.2 HMGB1 Ab ChIP

Nuclei were prepared as in bioChIP protocol and resuspended in sonication buffer consisting of 50 mM Hepes/KOH pH 7.5, 500 mM NaCl, 1 mM EDTA, 1% Triton X-100. Sonicated chromatin was precleared with fish skin gelatin and tRNA blocked protein-A Dynabeads (ThermoFisher, #10001D) beads for 1 h. Precleared lysate was incubated overnight with the HMGB1 antibody (Abcam, #ab18256) and subsequently with blocked protein-G Dynabeads (ThermoFisher, #10003D) beads for 3 h at 4 °C. Chromatin-bound beads were then washed twice with lysis buffer containing protease inhibitors at room temperature for 5 min followed by a single 5-min wash with 10 mM TRIS (pH 8.0), 250 mM LiCl, 0.5% NP-40, 0.5% Deoxycholate, and 1 mM EDTA. Beads were then transferred to a fresh tube and bound chromatin was eluted in elution buffer containing 1% SDS and 100 mM $NaHCO_3$ in two rounds of 20 min, rotating at room temperature. DNA extraction was performed as for bioChIP.

6.4.2.3 Alternative chromatin preparation protocol

One of the ESC HMGA1 and HMGA1 DBD replicate (designated HMGA1_c) was processed with a different chromatin isolation protocol.

Cross-linked cell pellets were resuspended in 50 mM Hepes-KOH (pH 7.5), 140 mM NaCl, 1 mM EDTA, 10% Glycerol, 0.5% NP-40, 0.25% Triton X-100
for 10 min on ice (membrane lysis). Nuclei were collected by centrifugation and resuspended in 10 mM Tris-HCl (pH 8.0), 1 mM EDTA, 200 mM NaCl, 0.5 mM EGTA for 10 min RT (removal of detergents). Nuclei were collected by centrifugation and resuspended in 10 mM Tris-HCl (pH 8.0), 1 mM EDTA, 0.1% Deoxycholate, 200 mM NaCl, 0.25% N-Lauroylsarcosin, 0.5 mM EGTA.
Crosslinked chromatin was subjected to sonication in a Bioruptor instrument (Diagenode). Triton X-100 to 1% final concentration was added before SAV-IP.

6.5 RNA-sequencing

For RNA-seq, two micrograms of total RNA from at least two independent cultures harvested on different days. RNA was isolated with the RNeasy mini kit (Qiagen) with on-column DNA digestion and was depleted from ribosomal RNA using the Ribo-Zero rRNA removal kit (Epicentre). Strand-specific RNA-seq libraries were prepared from rRNA-depleted samples using the ScriptSeq v2 protocol (Epicentre) following producer instructions. Up to 7 samples with different barcodes were mixed at equimolar ratios per pool. Sequencing was performed on an Illumina HiSeq 2500 machine (50 bp read length, single-end, according to Illumina standards).

6.6 Data analysis

All analysis was performed using R software unless specified.

6.6.1 ChIP-seq

6.6.1.1 FOXO1/3, SMAD3/4, SOX2

With reference to Section 7.1.2

ChIP-seq reads were mapped to the mm9 assembly of the mouse genome using bowtie (Langmead et al., 2009). Bowtie was run with parameters "-v 2 -a -m 100". For downstream analysis, reads were converted into the genomic ranges format (Lawrence et al., 2013), discarded if mapping to more than one genomic site and shifted according to the length of sonication fragments with chipseq R package (Sarkar et al., 2013).

Peaks were called with MACS-1.4.1 (Zhang et al., 2008) without input (--to-small --gsize=1870000000 --pvalue=1e-5 --tsize=50). Peak size was normalized over the peak summit and chromatin modifications (ESC H3K4me1: GSM769009; ESC H3K4me3: GSM769008; ESC H3K27ac: GSM1000099) were quantified over each TF protein-specific enriched regions. Mapped reads were normalized for library size (to the minimum) and a pseudo-count of 8 was added to overcome sampling noise. Annotation of known RefSeq transcripts was obtained from http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/refGene.txt.gz

With reference to Section 7.1.3

ChIP-seq and input samples were mapped to the mm9 assembly of the mouse genome using bowtie (Langmead et al., 2009) with parameters -v 2 –a –m 100, thus allowing up to 100 mappings per read. For quantification, alignments were given a weight of 1 divided by the number of mappings. Peaks were determined using MACS (Zhang et al., 2008) with parameters --tsize=50 --pvalue=1e-5 --lambdaset='1000,5000,10000' --nomodel --shiftsize=60 --gsize 1865500000.
After library-size normalization and shifting of reads by half the estimated fragment length (85nts), peak enrichments were determined as log2 enrichments over input, using a pseudo-count of 8. Motif finding at the peaks with strongest enrichments was performed with HOMER (Heinz et al., 2010) using the parameters *-size 1000 -N 5000 -S 5 -len 8,10 –nomotif* (as input to *findMotifsGenome.pl*) using known binding motifs from both Jaspar (Portales-Casamar et al., 2009) and a curated SELEX database (Jolma et al., 2013).

6.6.1.2 HMGB

ChIP-seq and input samples were mapped to the mm9 assembly of the mouse genome using the R package QuasR (Gaidatzis et al., 2015), which internally uses bowtie (Langmead et al., 2009). Bowtie was run with QuasR default parameters -m 1 --best --strata, thus allowing only for uniquely mapping reads. Further quantification was performed using the QuasR function *qCount* using half the estimated fragment lengths for the shift parameter and otherwise default parameters.
As the HMGB analysis presented here is of an exploratory nature, distinct methods were used for the HMGB analysis. These include varying ways of

estimating fragment lengths, varying window sizes to compare ChIP signals and varying definitions of mappable windows.

Judging from the data analysis so far, these details do not appear to lead to qualitative difference in the presented results. However, all results of this section should be considered preliminary at this point. Further details of particular analyses are indicated in the corresponding figure legends if deemed important.

6.6.1.3 HMGA

HMGA ChIP-seq and input samples were mapped in the same way as the HMGB samples. For the tiling window analysis, only windows in which at least 80% of the overlapping 50-mers were mappable using the alignment parameters described above were retained for further analysis. For quantification, the QuasR function *q*Count was used, adopting half the estimated fragment lengths as the shift parameter and otherwise default parameters. Fragment lengths were estimated by comparing read densities on the same and opposite strand of strongly bound CTCF sites and selecting the shift that minimized the root mean-square distance between the two profiles. This estimate appears more stable than fragment-length estimates at promoters due to the large number of positioned nucleosomes around bound CTCF sites and a more homogenous length of the nucleosome free region. Window counts of each sample were library-size normalized and enrichments over input were determined as *log2(IP + 8) – log2(Input + 8)*, where *IP* and *Input* are the corresponding counts per window and 8 is a pseudo-count that takes into account the increased noise levels at low read counts. Enrichment over the DBD mutant was defined as the difference in log2 enrichments over input (as defined above) between the wild-type and the corresponding DBD mutant sample.

6.6.2 Array data

For LaminA DamID, Dam ratio of the loess-quantile normalized data was downloaded from GEO (see accessions below). For the replication timing data, the wavelet-smoothed signal was downloaded from Encode (see accessions below). For both datasets, window levels were calculated by averaging the signal for each probe mapping to the respective window.

6.6.3 RNA-seq

RNA-seq reads were mapped to the mm9 assembly of the mouse genome using QuasR. Bowtie was run with mapping parameters "--trim5 3 -m 100 --best --strata". For multi-mapping reads, one randomly chosen alignment was used for quantification. The command used to perform the alignments in QuasR was proj<- qAlign("samplesRNA.txt", "BSgenome.Mmusculus.UCSC.mm9", alignmentParameter = "--trim5 3 -m 100 --best --strata"). Gene-level counts were determined using the UCSC knownGene table (Hsu et al., 2006)via the Rpackage *TxDb.Mmusculus.UCSC.mm9.knownGene* (Marc Carlson and Bioconductor Package Maintainer (2015). TxDb.Mmusculus.UCSC.mm9.knownGene: Annotation package for TxDb object(s)). As a stranded RNA-seq protocol was used, only reads on the same strand as the respective genes were counted (QuasR command qCount(proj, TxDb.Mmusculus.UCSC.mm9.knownGene, reportLevel="gene", orientation="same")). To determine significantly changing genes, voom as part of the limma R package was used (Law et al., 2014; Ritchie et al., 2015). Only genes that had a total count of at least 10 reads summed over all samples (after scaling gene counts of each sample to the smallest total gene count) were used as input for voom. To account for batch effects, a batch variable was included in the linear modelling which grouped samples when library preparation had been done on the same day and they had been sequenced on the same flow cell. Genes with an adjusted p-value < 0.01 (Benjamini-Hochberg) and an absolute fold-change of at least 2 were called as significantly changing.

For the quantification of repeat sequences, reads were remapped allowing only uniquely mapping reads (bowtie parameters "--trim5 3 -m 1 --best --strata"). Repeat Masker repeat annotation (Smit AFA, Hubley R, Green P. *RepeatMasker Open-3.0*. http://www.repeatmasker.org. 1996-2010) was downloaded from the UCSC genome browser (*http://genome.ucsc.edu/*, (Kent et al., 2002)). Repeat quantification was done on the level of repeat names using the qCount function of QuasR, counting only reads on the same strand as the annotated repeats (argument orientation = "same") and ignoring all repeats that overlap gene bodies (on the same or opposite strand of the corresponding gene). Repeats in gene bodies were excluded as their changes in read counts may be due to

expression changes of the corresponding genes. Significance of changes in repeat expression was calculated in the same way as in the case of gene expression.

6.6.4 PCA

Principal component analysis was run on samples indicated in Figure 7-11 with the R command "prcomp".

6.6.5 Accessions of published datasets used:

ESC H3K9me2: GSM1314605, GSM1314606, GSM1543602, GSM1543603

ESC H3K27me3: GSM632032, GSM632033, GSM632034

ESC Input: GSM671103

ESC PolII: GSM747547, GSM747548

ESC H3K4me1: GSM747542

ESC H3K4me2: GSM632035

ESC CTCF: GSM747534, GSM747535, GSM747536

ESC Sox2: GSM1050291, GSM1082341

ESC DNaseI: GSM1657364, GSM1657365

ESC and NP Bis-seq: GSM748786, GSM748787, GSM748788, GSM748789

ESC LaminA DamID: GSM1531435, GSM1531436

ESC Encode replication timing data was downloaded from:

https://www.encodeproject.org/files/ENCFF001JUP/@@download/ENCFF001JUP.bigWig;https://www.encodeproject.org/files/ENCFF001JUQ/@@download/ENCFF001JUQ.bigWig

RNA seq data for evaluation of HMG proteins expression: GSM687305, unpublished NP data (available on request) and GSM687306

6.7 Antibody used

Antigen	Supplier	Host	WB	ChIP
Lamin B1	Santa Cruz, C-20	goat	1:1000	
HMGB1	Abcam, ab18256	rabbit	1:1000	5 µgr / IP
HMGA1	Active Motif, 39615	rabbit	1:1000	
HMGA2	R&D Systems, AF3184	goat	1:1000	

Table 6-1 List of the antibody used in this study for WB and ChIP, with concentration used

6.8 Summary table of cell lines and data generated in this study

All cell lines are in the 159 background, which is a mixed 129Sv-C57Bl/6.

Primers used for cloning and cell lines are available upon request.

a

		GEX microarray (log2 values)			Stage		
		ES	NP d8	TN d6	Cloning	Cell line	ChIP-seq
bHLH	Neurod4	14	288	24	x	x	x
	Usf1	223	211	231	x		
	Neurog1	41	471	37	x		
SMAD	Smad3	175	668	258	x	x	x
	Smad4	686	1307	705	x	x	x
	Smad7	431	222	103	x		
Forkhead	Foxa1	47	90	45	x		
	Foxo1	293	206	70	x	x	x
	Foxo3	421	636	679	x	x	x
HMG-box	Sox2	1801	1026	233	x	x	x
	Sox4	197	1271	1538	x		
	Hmgb1	237	370	181	x	x	x
	Actb	*7342*	*6958*	*8944*			

b

Figure 6-4 a) Gene expression (Gex) data form GSE27114 for the selected TF and control gene. Bars in log2 space illustrate the difference in Gex; b) Summary of the dataset and cell-lines generated for HMGB and HMGA study.

7 Results

As of October 2016 the genome of Mus musculus encodes 1,476 proteins with an annotated DNA binding activity (source: curated database http://www.uniprot.org/). However for only a fraction of these proteins we have genome-wide location data. This is for two main reasons: first, only with the advent of next-generation sequencing of short reads (NGS) it became possible to precisely locate transcription factor binding sites (Valouev et al., 2008); second, many DNA-binders lack specific antibodies (Kidder et al., 2011). In an attempt to overcome these limitations, a former postdoc in the host laboratory devised a system, dubbed RAMBiO for recombinase-assisted mapping of biotin-tagged proteins (Baubec et al., 2013). This approach relies on site-specific integration of a desired expression construct, biotin tagging and chromatin location analysis by DNA sequencing. Using this streamlined approach he was able to map genome-wide methyl-binding domain (MBD) proteins and de novo DNA-methyltransferases (Baubec et al., 2013; 2015).

We started this doctoral project with the goal of mapping additional DNA-binding factors but we soon realized that we needed to prioritize our efforts. Knowing the importance of regulatory elements in transcription control, we decided to focus on DNA-binders that by previous studies had been implicated in promoter or enhancer function.

Ideal candidates were canonical transcription factors and we indeed chose a panel of 12 TFs for genomic location analysis. This served as a benchmark study for the second part of the project. There, we investigated the in vivo genomic preferences of High mobility group (HMG) proteins A and B.

7.1 Applying an antibody independent ChIP-sequencing paradigm to study the genomic location of TFs

The gold standard for determining in vivo TF binding sites is ChIP-sequencing of cross-linked chromatin, as this was the method of choice for the ENCODE and modENCODE consortia efforts (Odom, 2011). However as mentioned earlier, there is a need to expand the repertoire of antibodies that perform well in ChIP-sequencing experiments. An alternative strategy lies in engineering TFs to bear a protein tag and pulling down this invariant segment in order to enrich for bound

DNA. Biotin tagging and streptavidin (SAV) pull-downs represent a valuable tool for the molecular biologist. The dissociation constant (Kd) of $\approx 10^{-14}$ mol/L (Holmberg et al., 2005) allows stringent washes, which for experiments on cross-linked nuclei translates into dramatic reduction of the background signal.

In vivo biotinylation has been extensively used to study TF biology (de Boer et al., 2003; Wang et al., 2006; Zhao et al., 2014). With a ChIP-chip approach it was applied for the first time to investigate genomic location of pluripotency factors in mouse ESC (Kim et al., 2008). However the low-resolution of that study prevented the identification of individual TF binding motifs. More recently, several studies have reported successful application of a NGS adaptation of the technique for the study of additional TFs (Soler et al., 2011) (Soler et al., 2010) (Giraud et al., 2014).

We decided to benchmark RAMBiO performance for TF location studies with a panel of 12 TFs belonging to 4 different TF classes. TFs are separated in different families and classes according to the molecular structure of their DNA binding domain (DBD). The transcription factors we investigated possessed the Forkhead, SMAD, bHLH and HMG-box DBD domains. We included factors that are not normally expressed in ESC. Apart from factors for which no ChIP-seq data was available in ESC and also Sox2, a well-studied pluripotency factor, as an internal control (see Section 6.8 for a summary of the cell lines generated).

7.1.1 Testing feasibility, throughput and reproducibility of RAMBiO for TFs

A tagging approach such as RAMBiO has many points of strength when it comes to parallel in vivo assessment of TF binding. First, the presence of a tag makes the comparison between members of the same TF family more precise by abolishing the bias from different antibodies (Kim et al., 2009). Second, through stereotyped genomic integration comparable levels of transcription are achieved and thus endogenous protein abundance does not pose a detection limit. Third, ESC harboring the expression construct can be differentiated and binding re-assessed in a different chromatin environment. Other strengths are the possibility of testing functional mutants under the same conditions as WT proteins and the possibility to rescue KO cell lines in a reproducible manner.

In light of a potential scaling-up, it was important for us to get an estimate of the global throughput of the approach while assessing feasibility. This is why we decided to evaluate our approach on a panel of a dozen TFs belonging to different families and being differentially expressed in the neuronal differentiation paradigm we adopted in the laboratory (Lienert et al., 2011).

As can be seen in Figure 7-1a, out of the 12 proteins that we attempted to clone 82% were successfully retrieved from, either ESC, NPC or neurons. For the remaining proteins (potential causes can be low expression of factor or failed PCR) we needed to order synthetic constructs using annotated CDS (CIT materials and methods).

In RAMBiO, the parental ESC line harbors a negative selection cassette at the site of recombination. When the plasmid bearing the cDNA of the factor of interest is co-transfected with Cre recombinase, recombination can occur and recombinant clones loose the negative selection construct by dilution. The rate-limiting step of the protocol is the single cell cloning that ensues, which occurs during 10 days in selective media. After single cell cloning we evaluated protein expression by WB and in two thirds of the experiments we were able to isolate clones expressing the desired bioTF (Figure 7-1a).

Considering that we did not further investigate the reasons for the dropouts this is a remarkable result, which suggests that with few improvements (for example inducible expression or gateway cloning adaptation) the system could be readily utilized in even more comprehensive investigations.

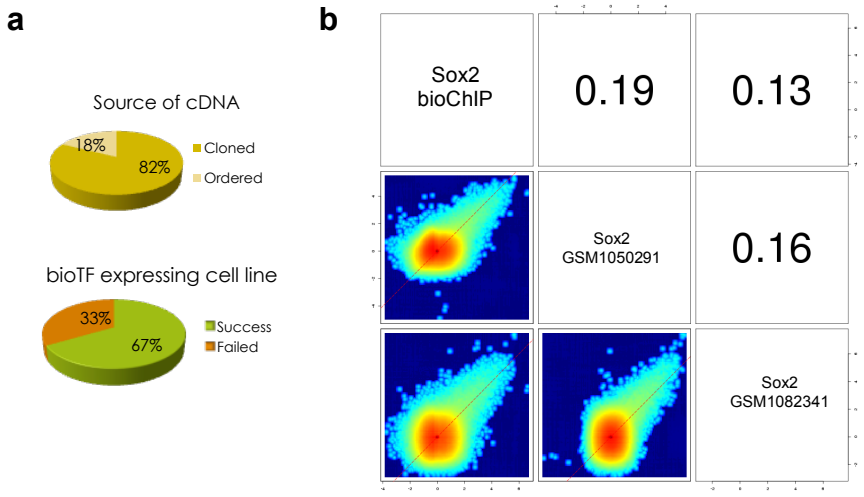


Figure 7-1 a) Pie-chart depicting source of cDNA and success rate in the generation of a panel of biotinylated TF expressing ESC; b) Scatterplots evaluating reproducibility of previously published Sox2 ChIP-sequencing data. Indicated are log2 enrichment values over input, in 500nts windows centered on predicted Sox2 binding sites. These were determined by scanning the genome with the Jaspar Sox2 motif and retaining sites with a log-odds score >=10 using uniform probabilities for each nucleotide as a background model. GEO entry number is indicated.

For those bioTF cell lines that we isolated we proceeded to perform bioChIP experiments. First, we checked whether TF data generated with bioChIP was in good agreement with antibody ChIP (Ab ChIP). We chose Sox2 as a proof of principle, since for this TF binding maps generated in ESC from the same genetic strain were available (Figure 7-1b). Globally one can see that Sox2 is enriched only in few genomic regions, therefore it is not surprising that correlation is low both between Ab ChIP replicates and with bioChIP sample. However, for regions that are enriched at least four-fold over input, bioChIP is in good agreement with the Ab ChIP data and shows variation with respect to the Ab ChIP data comparable to the variation between the two Ab datasets.

7.1.2 TF binding in relationship to chromatin and genomic features

TF tend to bind their different genomic targets in a manner that is associated to certain chromatin modifications (Cuellar-Partida et al., 2011). More specifically, binding of activating TF tend to occur in association with several types of histone marks (e.g mono- and tri-methylation of histone H3 lysine 4) (Heintzman et al., 2009) and with hypersensitivity to cleavage by DNase I (Neph et al., 2012). Even though we do not completely understand how individual chromatin marks affect TF binding and vice versa, this information can be used for other purposes. For example a variety of computational methods have been proposed that aim to improve our ability to identify active TFBSs on the basis of the DNA sequence plus epigenetic experimental assays (Mathelier and Wasserman, 2013; Whitington et al., 2011). Another useful observation is that the relative enrichment of different chromatin marks around bound sites can inform on the repressive/activating role of a TF (Ooi and Wood, 2007). Additionally it may also highlight potential preference for a specific subset of regulatory regions eg. enhancers or boundary sites (Ross-Innes et al., 2011).

With our set of experiments we were interested in determining the ESC binding maps of TF for which no data was available in this cell types, for example Foxo1-3. Since Neurod4 was not expressed in ES cells we were also interested in evaluating if and how, an exogenous TF could probe its canonical targets. Therefore we went on and looked for differences in the binding modality for the different TF factors.

Apart from Sox2, we looked at Foxo1-3 and Smad4 in more detail with respect to associated chromatin features (Figure 7-2a). Smad3 and Neurod4 samples gave a signal to noise ratio that only allowed very few regions to be reproducibly called as enriched. A plausible explanation for Smad3 could be lack of correct post translational modification (PTM) due to absence of upstream TGF signaling (Heldin et al., 1997). In the case of Neurod4, alternative possible scenarios are also the absence of binding partners or cell type specific modifiers.

Smad4 and Foxo3 proteins associated to both enhancer and promoter specific marks as can be seen in Figure 7-2a, looking at the chromatin signature around TF peaks. This data is also confirmed by looking at the distance of peak summits from the TSS genomic landmark (Figure 7-2b). For Sox2 we obtained a very different profile with a significant depletion of binding near promoter regions.

In conclusion, for the reduced set of factors that we studied, our data was able to unveil differences in peak-associated chromatin marks, which reflected factor specific skews in the type of regulatory region preferentially bound.

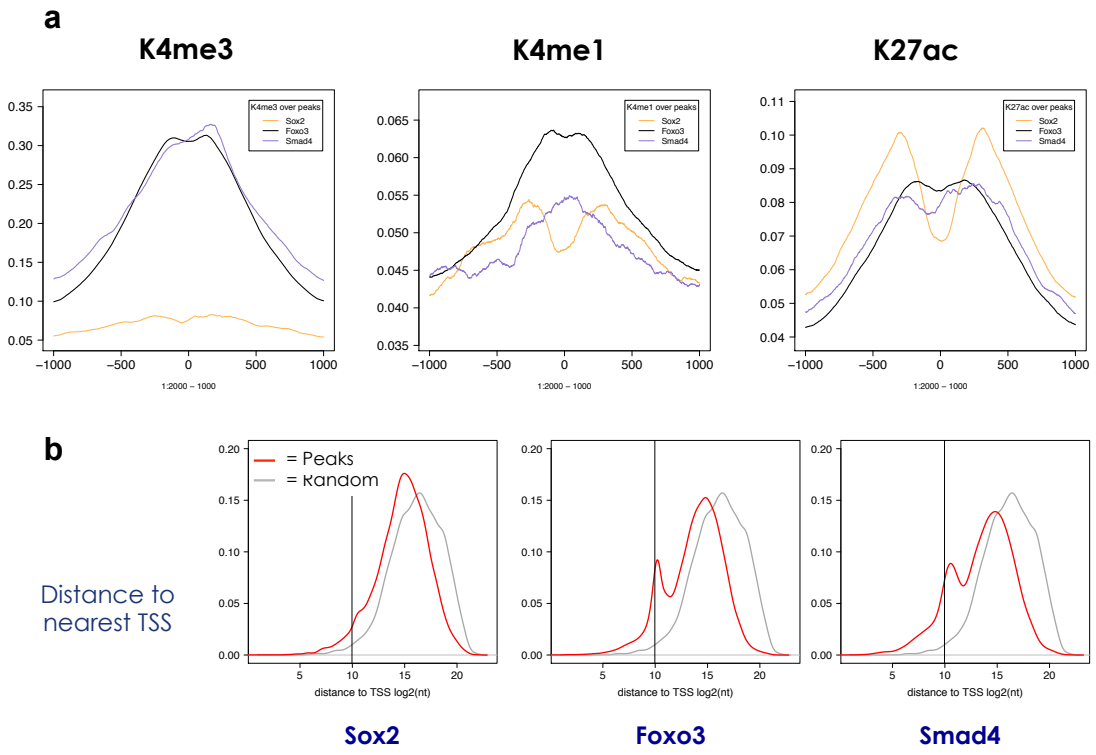


Figure 7-2 a) Metaprofiles of selected chromatin marks around peaks of different TFs. Sequences have been centered at peak summit and extended for the indicated length. A clear difference in histone mark distribution is apparent for Sox2; b) distance of the peak summit or of a random genomic nucleotide to nearest TSS. Grey line at approx. 1kb.

7.1.3 Accuracy considerations in the identification of TF motifs

TFs recognize short sequences (6–12 nucleotides long), often in a highly combinatorial fashion (Villar et al., 2014). At present, it is still impossible to accurately predict the genomic regions that will be bound by a certain TF in a specific tissue and at a specific developmental time. One of the reasons is that for only a minority of TFs the in vitro determined binding preference has been confirmed in vivo. In order to bridge this gap it is a priority to expand our repertoire of in vivo verified TF binding motifs.

To evaluate the power with which our technique could discover/recover in vivo binding preferences, we subjected bound regions to motif enrichment analysis, using known binding motifs from both Jaspar and a curated SELEX database (Jolma et al., 2013). Below is a summary of the motifs identified with the HOMER algorithm (Heinz et al., 2010). The heatmap shows known motif enrichments over peaks of the indicated TF (Figure 7-3). Fox motifs were enriched under the

respective bioChIP experiments, and so was Sox2 and, to a lesser extent, the Smad motif in the Smad4 peaks. Analysis was replicated for three different subgroups of peaks, divided for their binding strength. For both REST and Fox1/3, the motif enrichments decrease with decreasing binding strength, indicating the quantitative nature of the ChIP signal. These results underscore the high sensitivity and reproducibility of bioChIP: RAMBiO results captured known TF preferences, thus putting a high degree of confidence to data coming from proteins with undetermined motifs. This finding is important for the evaluation of the results for HMGA and HMGB that will be discussed in the next sections.

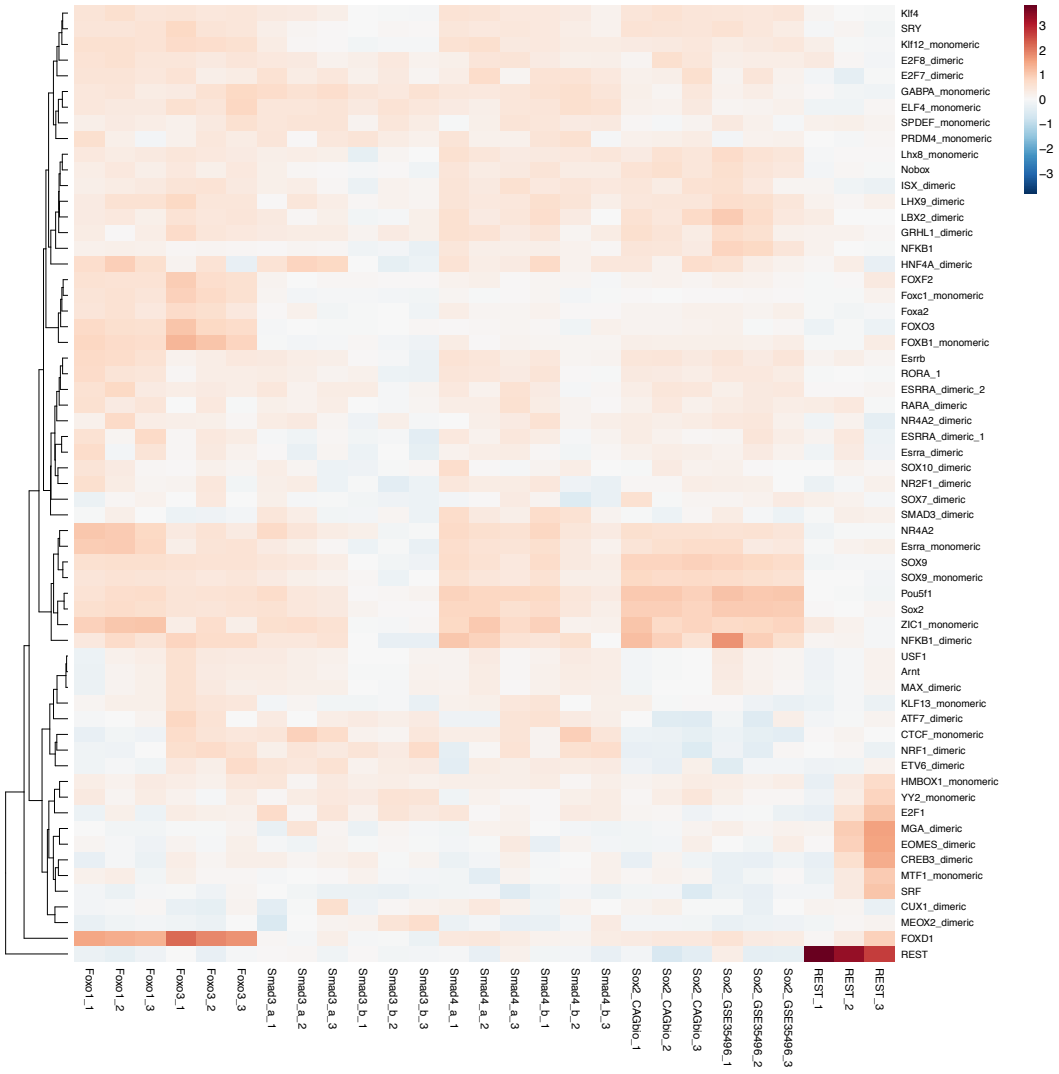


Figure 7-3 Heatmap of the HOMER (Heinz et al., 2010) results for the TF panel and two antibody based ChIP-seq experiments (REST GSE27148 and Sox2_GSE35496). Each sample data is divided in three sets of regions: first 1,000 peaks, second and third 1,000 peaks are assigned respectively number 1-2-3. Color indicates log2 enrichment.

7.2 Genome-wide location analysis of HMGB proteins

The next class of protein that we focused on is HMGB. This is a conserved family of proteins whose members are found throughout evolution from yeast, to plants to sponges and animals (Bianchi and Agresti, 2005). In the mouse there are 4 members that fall into this category of small nuclear proteins. All of them have a tandem HMG-box DBD domain and all but one have a C-terminal unstructured acidic tail. Three of them are expressed throughout the body-plan, however in the hematopoietic system and in embryonic tissues they are expressed the highest. HMGB4, the one that lacks the acidic tail, is expressed only in round spermatids at the time of histone to protamine exchange (Catena et al., 2009).

These proteins have been implicated in many aspects of nuclear biology, from enhanceosome stabilization to histone chaperoning, to DNA-damage and recombination (Stros, 2010).

However at the time we started our study no genome-wide location data was available for any of the factors that could generalize or confirm in live cells the observations made in vitro.

We cloned therefore all members of the HMGB family and generated bioChIP data in both ESC and NPC.

In ESC only Hmgb1 and Hmgb2 are expressed. Hmgb1 expression by RNA-seq quantification is difficult because the signal is diluted across the cDNA of the numerous pseudogenes (Figure 7-4a). Its expression can be better inferred by looking at the RNApol2 quantification around its promoter (Figure 7-4b). We first checked protein production and correct subcellular targeting of the bioHMGB proteins by contrasting to HMGB1.

We expressed HMGB proteins to a level comparable to endogenous HMGB1, as assessed by WB with a specific antibody (Figure 7-4c). For the other HMGB members we checked protein expression by SAV blotting.

With SAV staining we checked the subcellular localization of bioHMGB1 (representative of bioHMGB2-3, as from available immunofluorescence data (see Materials and methods section) and bioHMGB4 proteins (Figure 7-4d). We could see that HMGB4 diffuses freely in interphase nuclei, whereas HMGB1 diffuse freely but is also enriched at nucleoli. This was in good agreement with known

subcellular localization of HMGB1 in fixed cells (Pallier et al., 2003) and we therefore considered our bioHMG proteins as functional.

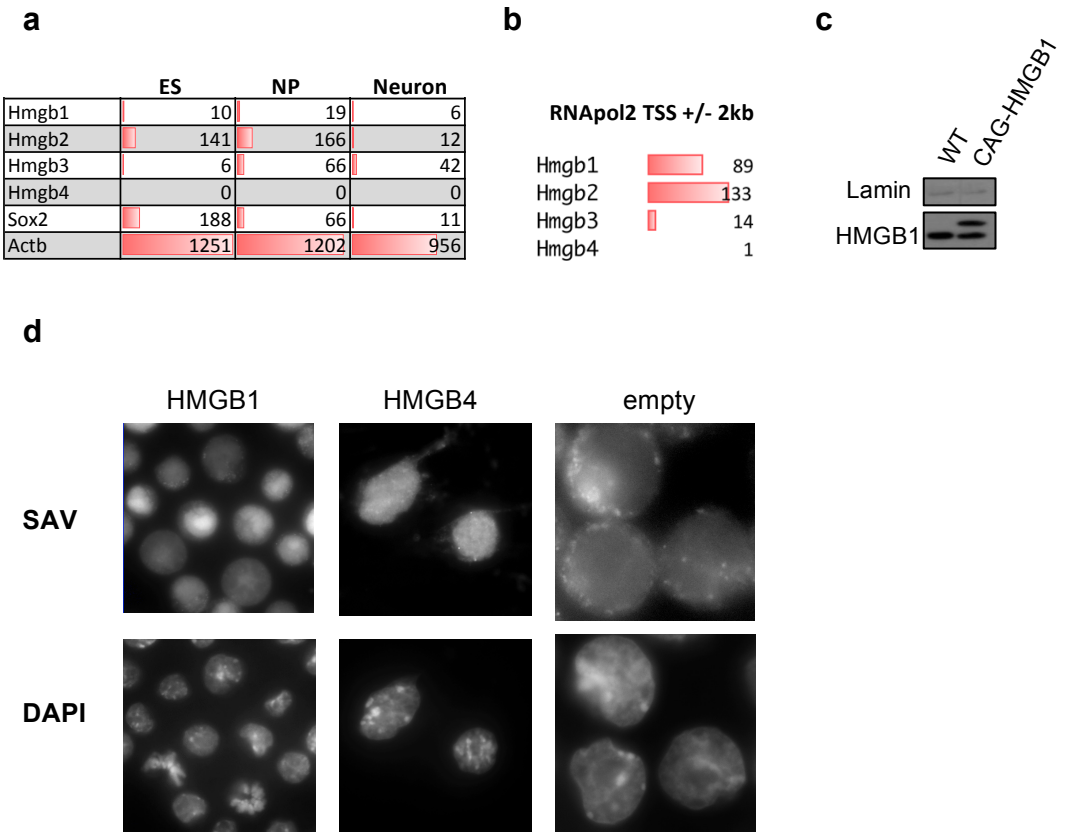


Figure 7-4 a) Unique RKPM for ESC, neuronal precursor and TN over HMGB genes. Only HMGB1-2 are expressed in ESC. b) Quantification of RNApol2 ChIP sequencing over a 4kb region centered around TSS of the indicated genes; c) WB of WT and bioHMGB1 expressing cell line with HMGB1 antibody; d) SAV-fluorophore fixed cells imaging of HMGB1, HMGB4 and parental cell lines. HMGB1 is released from mitotic chromatin. In the parental cell line, residual staining is observed at mitochondria due to endogenously biotinylated carboxilases.

7.2.1 Global characteristics of HMGB binding

Having assessed that HMGB can be expressed and correctly localize in nuclei, we proceeded to carry out bioChIP for the HMGB samples. For all samples DNA IP efficiency was low (below 1/10'000, from 5pg/cell maximal theoretical calculation and measured from input DNA quantification) but library preparation was possible in all cases. For HMGB2 and HMGB3 we saw a very similar binding profile to HMGB1 (Figure 7-5a), and for this reason HMGB1 related finding can generally be extended to HMGB2-3 and vice versa.

We noticed by visual inspection in the genome browser that for all proteins enriched regions coincided with open chromatin as assessed by DNaseI, with a larger dynamic range for HMGB4 (Figure 7-5b).

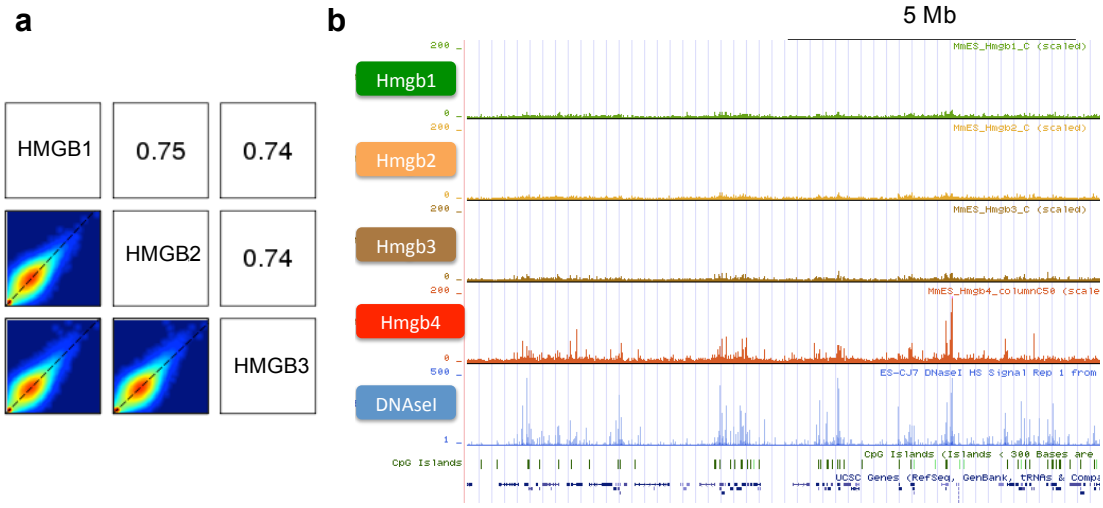


Figure 7-5 a) Scatter plots and Pearson's correlation coefficients of ESC bioChIP reads for HMGB1-2-3 averaged over 500 bp tiling windows on Chr1. Axes of the individual scatterplots range form library size normalized log2 read counts 3 to 8; b) UCSC browser snapshot of library size normalized read counts for the indicated HMGB proteins in ESC at Chr3 centered at bp 35M. In blue is a DNaseI track from ENCODE consortium. CpG-island track and UCSC gene annotation track are also displayed.

7.2.2 Dissecting binding to open chromatin: GFP bioChIP and role of DBD

The finding of HMGB proteins accumulating at open regions matched previous and concurrent observations for HMGB2 and HMGN (Cuddapah et al., 2011; Deng et al., 2013; Redmond et al., 2014; Zhang et al., 2016). However since some antibodies are prone to pull-down open regions of DNA, at least in Drosophila (Jain et al., 2015), we reasoned similarity with this type of binding should be taken with caution. We wondered whether the enrichment at active regulatory regions for HMGB2 and HMGNs could be due to poor antibody specificity or random unspecific contact of these proteins with open regions. Regarding this last point it is known from a study in yeast that indeed GFP protein accumulates at open regions of DNA (Teytelman et al., 2013). We therefore generated a cell line expressing monomeric GFP and carried out bioChIP experiments with both standard ChIP crosslinking conditions and 1h at 4°C (as used by Redmond et al.

2014). Genomic profiling of GFP indeed showed tendency to make unspecific contacts with DNA at open regions also in the mouse genome (Figure 7-6a). Importantly GFP relationship with hypersensitivity to DNaseI (DHS) was comparable to that of HMGB1-2 and this was true for HMGB1 also by standard antibody ChIP-seq (Figure 7-6a, see figure legend).

Given the similarity with GFP, we then wondered whether HMGB1 binding to DNA was mediated through its binding domains or via random contacts. To this end we expressed a mutant version of HMGB1, mutated at three key residues in the two HMG-box DBD that render the protein more diffusible in vivo and less bound to DNA substrates in vitro (Agresti et al., 2005; Jung and Lippard, 2003). Looking at the correlation between binding and accessibility, we observed a similar pattern for the mutant protein as for the WT (Figure 7-6a). We confirmed this observation by directly comparing binding of the mutant and WT in Figure 7-6b. Indeed HMGB1 mutant binding is indistinguishable from two HMGB1 replicates (Pearson's correlation 0.54 vs 0.55 when looking at windows). This analysis suggests that in mouse ESC HMGB1 DBDs are not required for the observed interaction with DNA.

This might mean that in this cell type either HMGB proteins are kept in an inert state by post-translational modifications, are sequestered by a yet undetermined inhibitor or do not interact with genomic DNA.

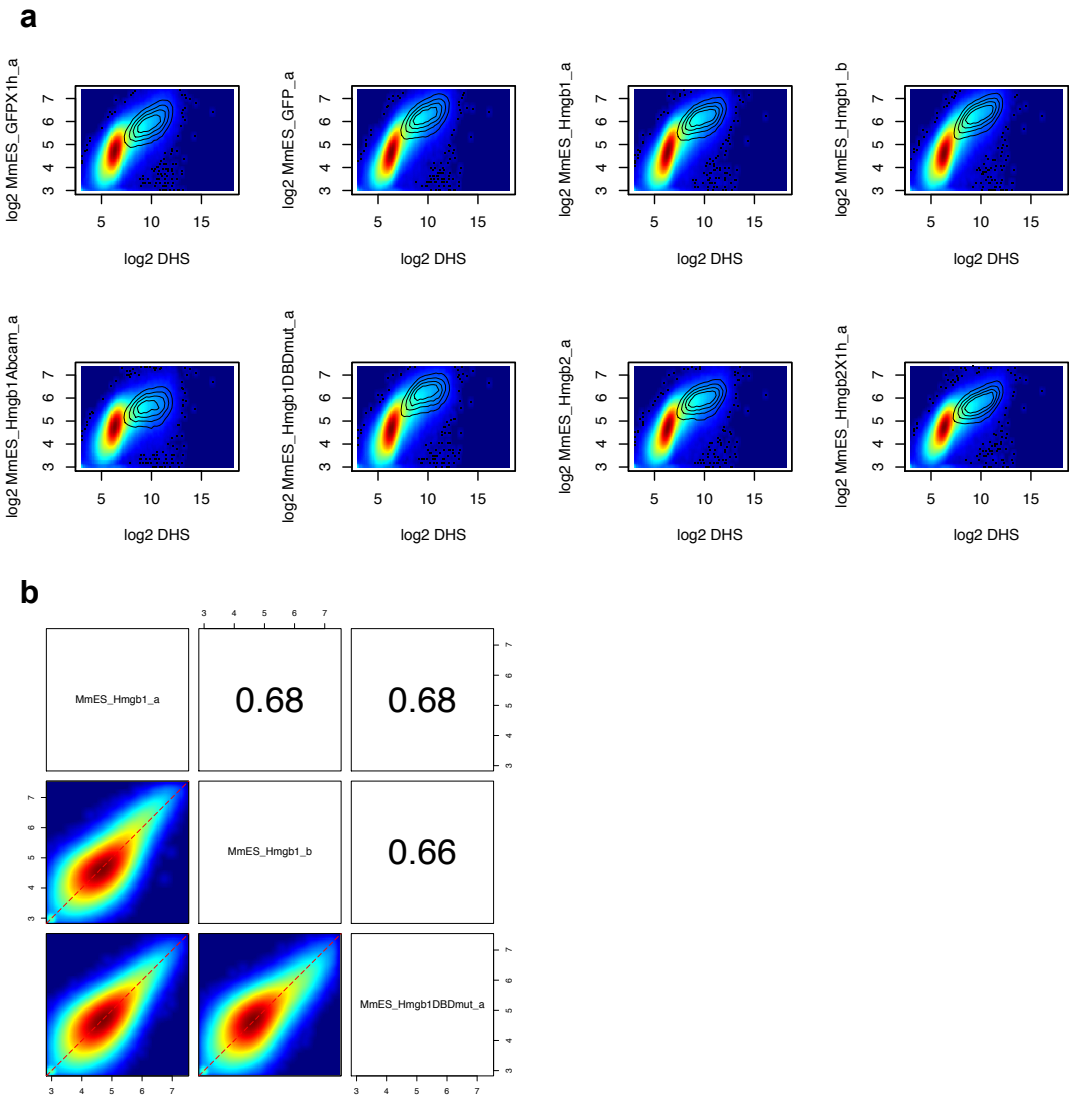


Figure 7-6 a) Scatterplot of read count versus DNaseI cut frequencies for various replicates including the 2h at 4° C crosslinking. MmES_HMGB1Abcam_a, antibody ChIP-seq in the parental cell line that in our hands yielded a very poor signal to noise ratio. Nevertheless allowed to confirm for the endogenous HMGB1 protein that regions with higher HMGB1 binding were also CGI regions with accessibility higher than average. Analysis over 500 bp tiling windows. Contour lines highlight the density of regions overlapping with UCSC CGI; b) Pearson´s correlations and scatter plots between log2 transformed reads averaged over 1kb tiling windows for two HMGB1 replicates and a HMGB1 mutant mutated at three key residues in the two HMG-box DBD that render the protein more diffusible in vivo and less bound to DNA substrates in vitro (Agresti et al., 2005; Jung and Lippard, 2003).

7.2.3 Investigating residual HMGB4 enrichments after GFP signal subtraction

As an additional way to examine HMGB1 affinity and specificity for DNA in vivo, we subtracted GFP signal from that of HMGB. If we were to observe no enrichment the conclusion would be that HMGB binding to DNA is highly similar to GFP. For HMGB1, indeed no regions are found reproducibly enriched over GFP (Figure 7-7a). For HMGB4 we saw however residual correlation with openness in

the two HMGB4 replicates and it also seemed that the highest binding occurred at DHS site that are not CGI (Figure 7-7a). This could be due to specific HMG4 protein-protein interactions (PPI), smaller size/absence of an auto-inhibitory domain (Sheflin et al., 1993) or sequence-specific recognition of a DNA motif.

In order to test this latter hypothesis we generated a truncated version of HMGB1 lacking its acidic tail that was of a comparable length to HMGB4. However binding of this protein to the genome was almost identical to its full-length form (Figure 7-7b). These results indicate that the acidic tail of HMGB proteins is not involved in inhibiting HMGB1 binding to DNA.

An alternative scenario for explaining the HMGB4 over GFP enrichments would be the recognition of a specific subset of regulatory regions due to base pair recognition. If this were true, one would expect the same sequence to be recognized in different cell types. In order to address this point we differentiated ESC towards neurons and harvested NPC to perform bioChIP experiments. At this stage cells have undergone a transcriptional remodeling and there is a reshaping of openness (Domcke et al., 2015) mainly occurs at non CGI regions (Stadler et al., 2011). We saw remodeling of HMGB4 binding in correlation with changes in DNA methylation (Figure 7-7c) We then used HOMER (Heinz et al., 2010) for motif enrichment analysis. The motifs that we found enriched under HMGB4 highly bound regions in the two different cell types are however largely different (Figure 7-7d). Besides, the top motifs that are called in each experiment belong to known master TFs for that cell-type or are called from a ChIP-seq experiment performed in the same cell-type. Therefore it seems that the DBDs of HMGB4 does not have an intrinsic DNA sequence preference.

Altogether, we show that HMGB4 tends to occupy accessible sites more frequently than GFP in a sequence independent manner and for reasons that are unrelated to the absence of an acidic tail.

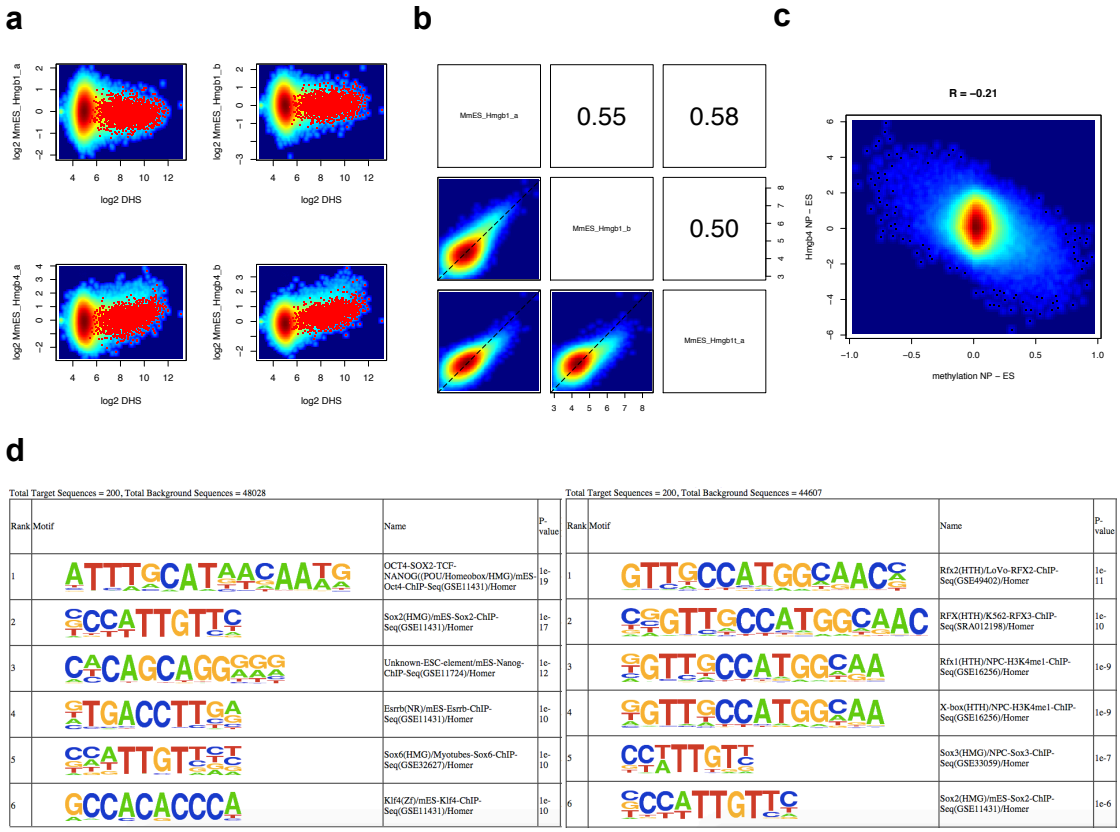


Figure 7-7 a) Scatter plot of the GFP-subtracted signal from two replicates of HMGB1 and HMGB4 vs. DNaseI cut frequency (500 bp tiling windows Chr1); b) Pearson´s correlations and scatter plots between log2 transformed reads averaged over 500 bp tiling windows on Chr1 for two HMGB1 replicates and one HMGB1 mutant lacking the acidic tail; c) Scatter plot of HMGB4 bioChIP and % methylation (1 indicating 100%) from ES to NP (methylation represents difference in average methylation in 500nt tiling windows on Chr1); d) Motif enrichment results for top 200 HMGB4 peaks in ESC (left) and NPC (right). Some of the NPC motifs resemble a TATA-box consensus. Motif finding was performed using HOMER (Heinz et al., 2010) with parameters -size 500 –nomotif –mknown using the vertebrate weight matrices that are part of the HOMER software. Peak size was adjusted for both samples to a standard length centered at peak summit.

7.2.4 Genetic rescue of isogenic Hmgb1 KO cell line and further assessment of HMGB1 functionality

An important point for the mapping approach adopted in this thesis is to assess whether tagged proteins are correctly folded and functioning after biotin conjugation.

In order to address this point we first generated Hmgb1 KO, with the aim of rescuing possible phenotypes through biotin tagged add-backs. However in our cellular system no overt differences could be observed between WT and Hmgb1 CRISPR Cas9 KO at the transcriptional level (Figure 7-8a). This was the first transcriptomics characterization of Hmgb1 KO to our knowledge. A previous study showed changes in histones and total RNA content upon HMGB1 KD in

MEF (Celona et al., 2011), however that transcriptional change could represent transient phenomenon that is not fixed upon sustained HMGB1 reduction. Since practically no genes (HMGB1 and pseudogenes excluded) were changing significantly we did not have the opportunity to test the effect of bioHMGB1 reintroduction.

Therefore we adopted an alternative strategy to reduce the probability of having been observing the properties of a dysfunctional protein. In order to extend the validity of our findings we decided to place the tag to the C-terminus of the HMGB1 cDNA and repeat bioChIP experiments. As can be seen in Figure 7-8b the two differentially tagged proteins bind the genome and enriched regions in a similar way. Even though this result does not rule out completely the possibility of a tag-induced functional impairment, it is unlikely that in both scenarios (N and C-tagging) biotin affected binding in a similar way. Finally, in previous studies from another group there are proofs of functionality for HMGB1 C-terminally fused to GFP (Agresti et al., 2005), which is approximately 10 times larger than a biotin tag.

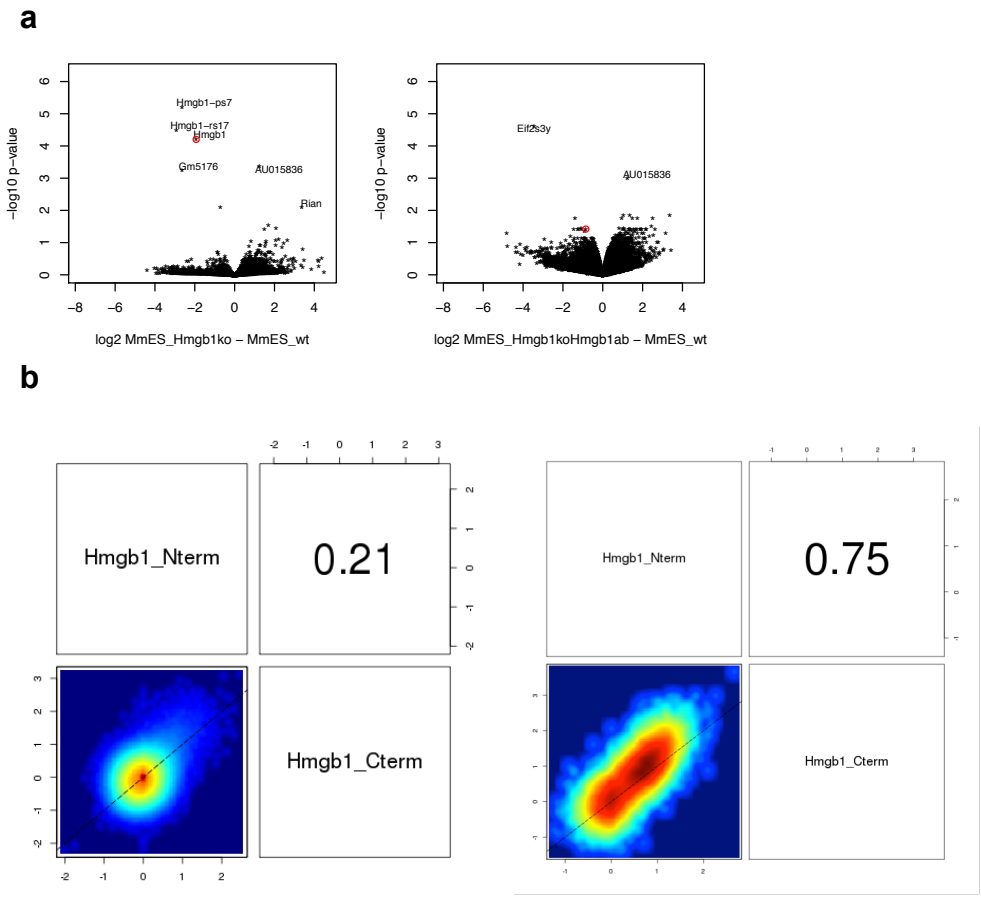


Figure 7-8 a) Volcano plot of the Hmgb1 and Hmgb1 KO with HMGB1 add-back vs. WT. Gene name is indicated for adjusted p-value < 0.01 and expression change of at least 2 fold. The Hmgb1 gene is shown in red. Hmgb1 pseudogenes calls are likely partially alignment artifacts. Analysis min 2 replicates, max. alignment repetition = 100, Limma Voom R package; b) Pearson´s correlation and scatterplot of log2 enrichment over input signal for C-terminus and N-terminus tagged HMGB1 constructs. Left, 1kb tiling windows Chr1. Right, promoter regions genome-wide.

7.2.5 Conclusion and future perspective

Overall, our data indicates limited binding to genomic DNA for HMGB members. However, given the high similarity in binding with inert and DBD mutated proteins, a more thorough ascertainment of protein functionality is required. In light of the interesting results we obtained with HMGA proteins we nevertheless decided to focus on the latter, rather than investing on additional control experiments for HMGB proteins, which would be required for a clear description of HMGB binding.

7.3 Genome-wide location analysis of HMGA proteins

HMGA proteins are small nuclear proteins robustly expressed during embryonic development and in fast replicating cells e.g. in certain hematopoietic lineages (see Introduction). They are also found misregulated and/or truncated in a number of cancers (Benecke et al., 2015; Peter et al., 2016; Wood et al., 2000).

Originally, HMGA1 was described and cloned because of its binding to a human major satellite sequence (Strauss and Varshavsky, 1984). Subsequent in vitro studies characterized the preference of HMGA1 DBD for AT-rich DNA, and the domain was named 'AT-hook domain' after this discovery (Reeves and Nissen, 1990; 1993). However surprisingly little is known about HMGA1-2 binding in vivo. Location of HMGA proteins has been investigated so far only in one human cell type (Winter et al., 2011). In that study the authors show AT preference in vivo however the claim relies on a very low throughput ChIP approach. In brief, DNA extracted from HMGA2 ChIP was cloned in bacterial plasmids and 49 colonies were then genotyped by Sanger sequencing. This is a huge underrepresentation of the genome.

To investigate the detailed genomic location of HMGA proteins we therefore decided to utilize our RAMBiO approach.

We started by designing recombination constructs for isoform A of HMGA1 (HMGA1 from now on) and HMGA2 proteins. After transfection, we isolated and characterized single cell clones of mouse embryonic stem cell (ESC) harboring the integrated construct. As can be seen in Figure 7-9a bioHMGA1 is expressed to comparable levels to endogenous HMGA1. HMGA2 endogenous protein on the contrary is not expressed in ESC, in agreement with RNA-seq expression profiling (Figure 7-9a).

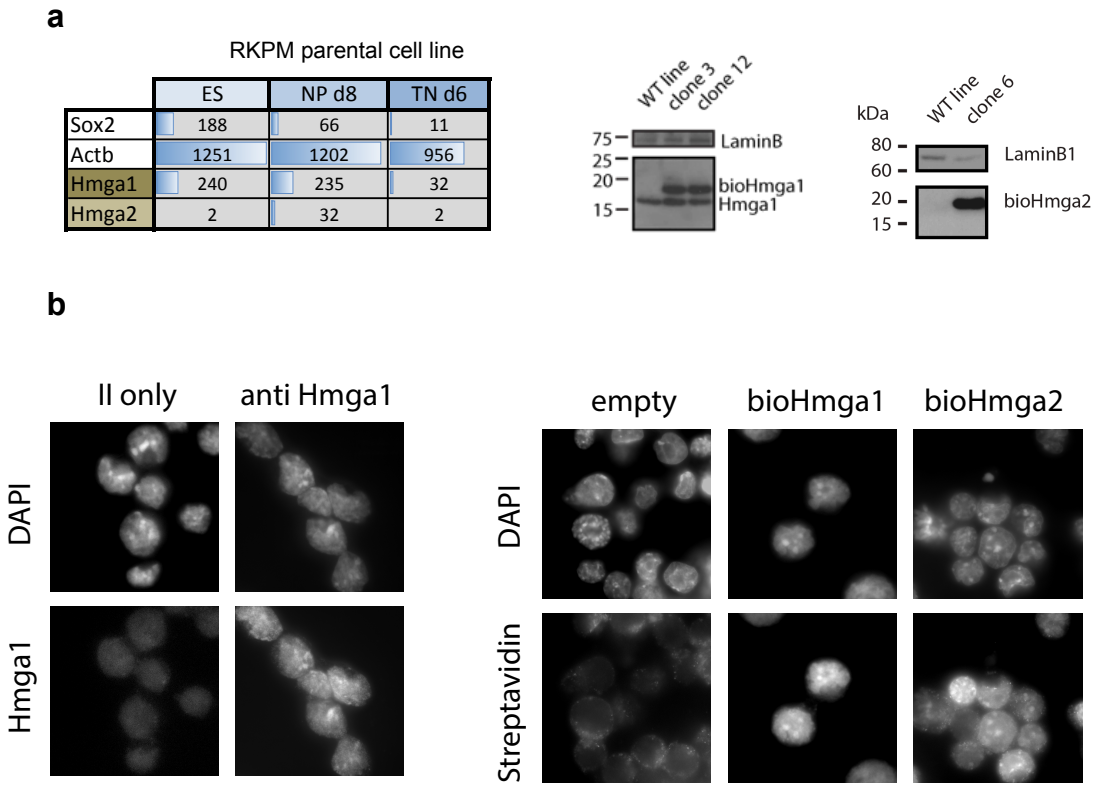


Figure 7-9 a) Left, tables shows RPKM for HMGA1, HMGA2 and two control genes. HMGA1 has one pseudogene, therefore mapping was done allowing multiple alignments and therefore reported values are an underestimation of the expression levels; Middle, WB of two clones expressing HMGA1 and parental cell line, blotted with anti HMGA1 Ab. Right, WB of one clone expressing HMGA2 and parental cell line, blotted with anti HMGA2 Ab; b) Left, subcellular localization of HMGA1 by antibody immunofluorescence. Right, subcellular localization of HMGA1-2 by conjugated SAV-fluorophore microscopy.

We checked the subcellular localization of both proteins by immunofluorescence and in line with previous observations (Disney et al., 1989; Henriksen et al., 2010) we observed enrichment at DAPI dense foci (Figure 7-9b). Given the correct localization of the biotinylated proteins we proceeded to perform bioChIP experiments.

After pull-down a considerable amount of DNA was retrieved (up to 1/500 of the DNA subjected to IP, see section 7.2.1 for a comparison) pointing to a high intrinsic affinity of HMGA1-2 for DNA.

7.3.1 A genome browser view of HMGA1-2 binding

We first evaluated the distribution of aligned reads by visual inspection in a genome browser: reads were evenly distributed over cis-regulatory regions, genes and intergenic regions in a manner reminiscent of input DNA. This initial observation signifies that HMGA1-2 contact DNA throughout the genome with similar strengths.

To test whether indeed there was no enrichment of binding after SAV-mediated pull-down, for each experiment we sequenced corresponding input DNA. This DNA comes from direct extraction of the sonicated chromatin material prior to IP and represents the IP substrate. Of note, during library preparation for NGS we applied the same number of PCR cycles per amplification as the IP fraction. This is an important detail since it is known that sequencing data suffers of a complex systematic bias and PCR is the most important cause of this bias (Benjamini and Speed, 2012).

As we did not detect punctuated binding upon visual inspection, we chose to initially investigate ChIP enrichments over input on relatively broad regions of 10kb using tiling windows along the genome. We opted for this value reasoning that if we would not observe enrichment averaging over such large regions, then HMGA1-2 binding truly showed no genomic preference at all. However enrichment over input analysis highlighted a localization pattern that was consistent between replicates and shared between HMGA1-2 (Figure 7-10 in purple and green).

In order to test if the broad binding pattern was specific, it was crucial to also generate ChIP-seq data for a DNA binding domain mutant (DBDmut) for both HMGA1 and HMGA2. These proteins harbor R>C point mutations at key residues of their DBDs which were previously shown to be important for DNA binding in vivo (Harrer, 2004). Interestingly, already after performing SAV-precipitation it was obvious that proteins' affinity to DNA was compromised judging from the little DNA recovered (same efficiency as HMGB proteins, see section 7.2.1). As can be observed by looking at the lower tracks in Figure 7-10, DBD mutants were more frequently found at regions of lower HMGA enrichment.

To assess the nature of this DBD mutant enrichment, we also plotted GFP bioChIP results, using the same enrichment over input metrics. Reassuringly,

HMGA1-2 DBDmut tracks showed extensive similarity in terms of regions enriched with GFP (Figure 7-10 yellow track). The intensity of this residual, unspecific binding seemed stronger for HMGA1-2 mutant proteins as compared to GFP, possibly due to their smaller size (less than half: 106-8 vs. 238 aa) and thus higher diffusion coefficient.

Taken together, this data clearly shows that WT proteins are binding to DNA in a DBD dependent manner.

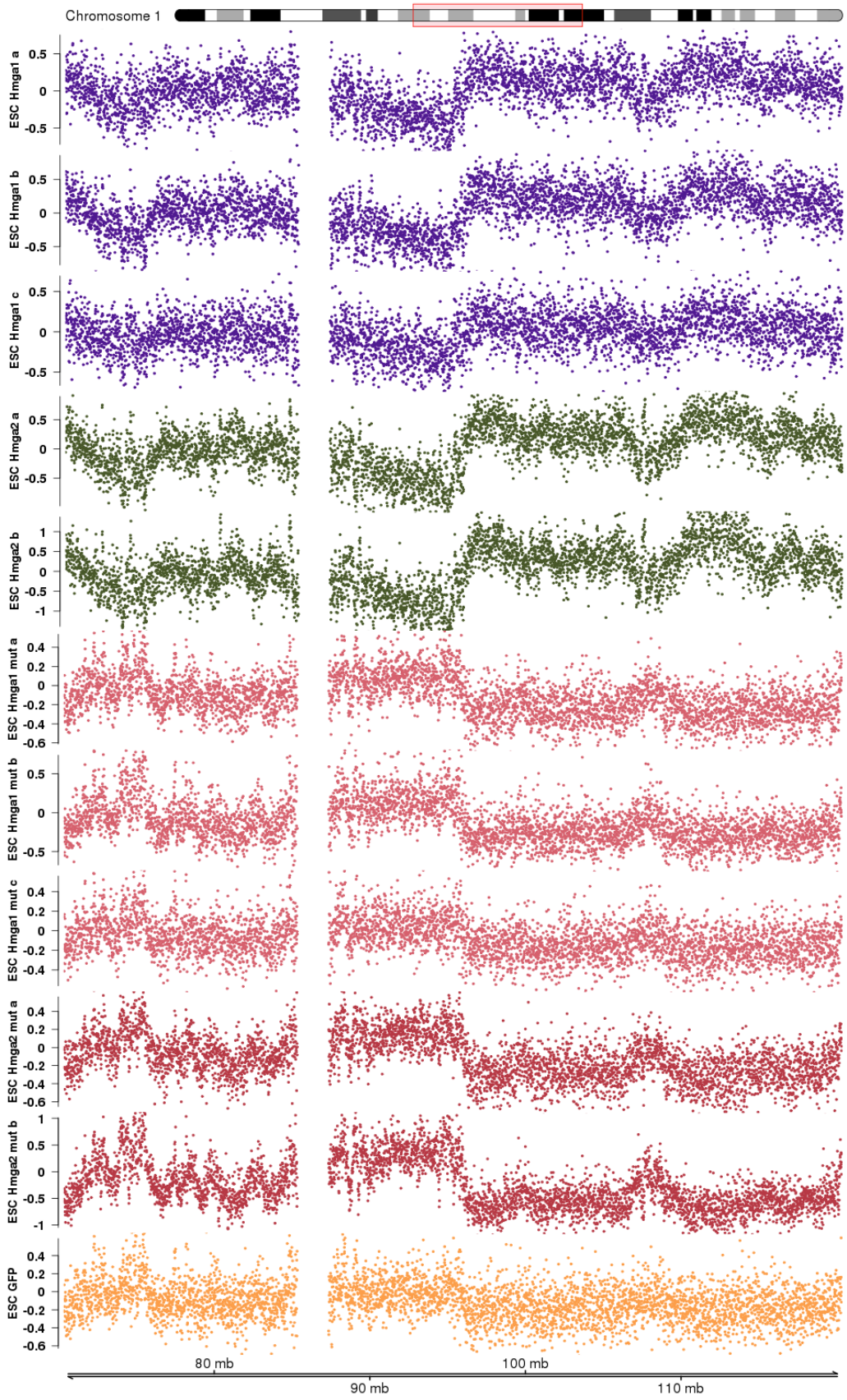


Figure 7-10 Average signal intensity of log2 enrichment over input for the depicted samples over a fraction of Chr1. For better readability top and bottom 1% of data range not shown Data for replicate C of HMGA1 and HMGA1 mutant was obtained from a different ChIP protocol preparation to highlight

(continues)

robustness (see Materials and methods). Each dot represents the log2 enrichment of IP over input in a window of size 10kb.

7.3.2 Principal component analysis to uncover binding determinants

Having assessed the bona-fide global affinity for DNA we set out to determine potential features that are responsible for the observed binding by means of principal component analysis (PCA). PCA is a mathematical algorithm that enables dimensionality reduction of large datasets. It does so by identifying directions, called principal components (PC), along which the variation in the data is maximal (Ringnér, 2008).

The analysis was carried on the GFP sample, two replicates each of the DBD mutant and the WT proteins, using log2 enrichments of IP over input in tiling windows of 1kb along the genome. Interestingly, in our dataset a single PC could explain almost 40% of the total variance in the data (Figure 7-11a).

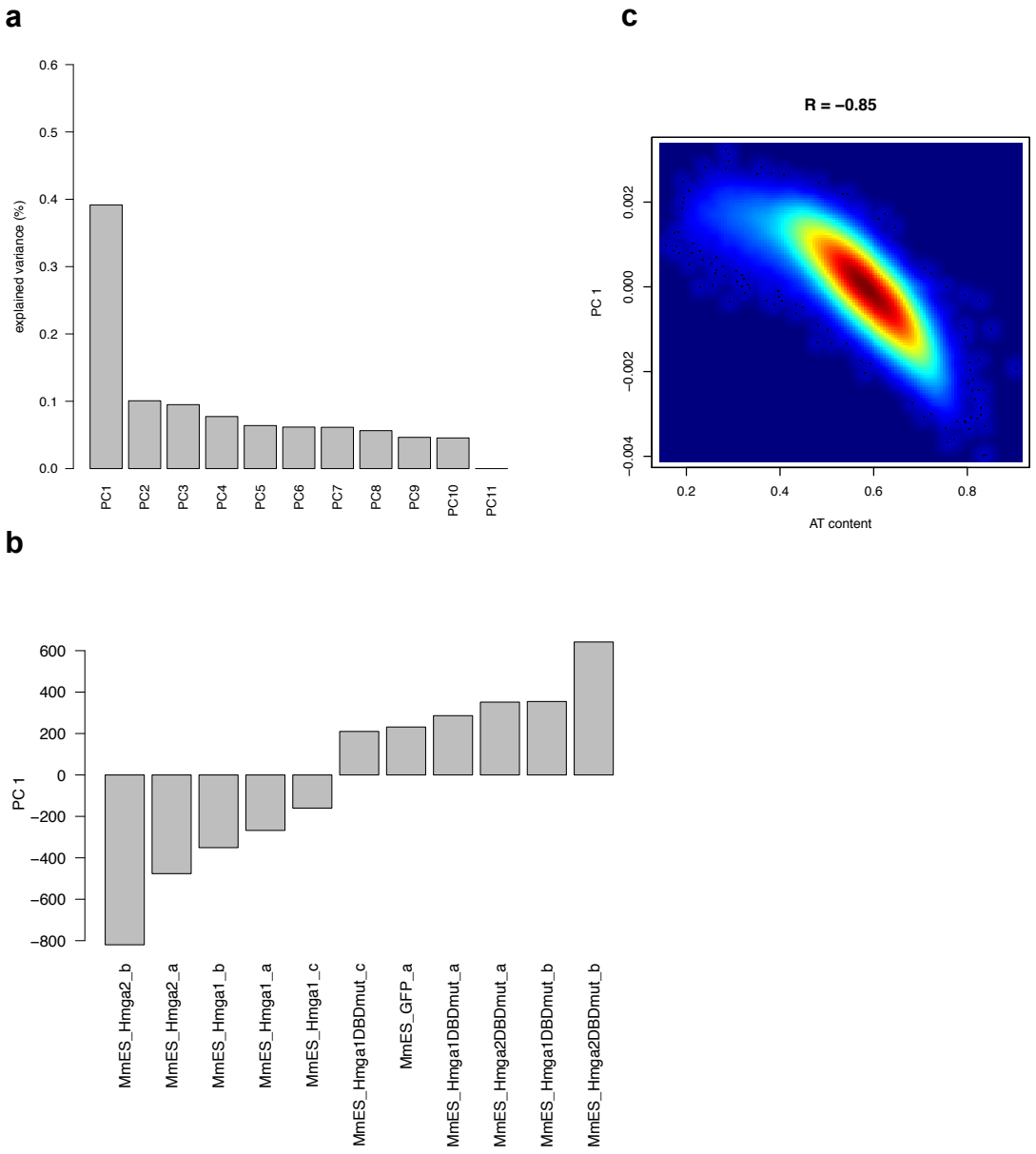


Figure 7-11 Principal component analysis of a dataset comprising DBD mutated and WT bioHMGA samples in ES cells. a) Fraction of the total variance explained by each PC = Principal Component. b) PC1 coefficients for each of the indicated sample´s enrichment values; c) scatterplot and Pearson correlation of PC1 with AT-content.

The coefficients with respect to PC1 gave us indeed a clear separation between HMGA1-2 proteins and GFP or DBD mutant proteins (Figure 7-11b).

In an attempt to link the first PCs to physical variables we contrasted them with marks of chromatin states, genomic features and primary sequence metrics. This revealed that PC1 was highly correlated to AT-content, calculated as percentage

of A or T in 1kb tiling genomic windows (Figure 7-11c) (note that the sign of the correlation is arbitrary).

7.3.3 Assessment of AT-content dependence for HMGA1 and HMGA2

After having established that AT-content explained a large fraction of the variance in our data, we asked to which extent the actual HMGA1 and HMGA2 signal scaled with AT-content. In other words we wanted to know for example whether HMGA response to AT-content was linear or whether it resembled a sigmoidal curve, pointing to a threshold effect.

To this end, we divided the genome in consecutive windows of 1kb, as was done for PCA, and looked at enrichment values over AT (Figure 7-12a). This analysis for AT-dependence showed positive correlation between AT-content and HMGA1-2 enrichments for all samples. Importantly no enrichment was observed at CGI (contour plots) in sharp contrast to the DBDmut samples. For DBD mutants we replicated the results observed previously for GFP (see Figure 7-6) where enrichment over input is highest at CGI. The data also shows that the dependence to AT was roughly linear for both HMGA1 and HMGA2 outside of CGI. This is of interest since few DNA binding proteins with DNA recognition of low information-content have been investigated so far. A notable example are MBD proteins, for which (with the exception of Mbd3) a linear correlation was also observed, in this case between binding and methylation density (Baubec et al., 2013).

Next, since different tissue expression between HMGA1-2 may imply non-redundancy, we examined potential differences in the strength of the AT-dependence. After input normalization, we compared AT-dependence by subtracting from the log2 enrichment values (over input) of each protein the enrichment (over input) of the respective DBD mutant. By doing so we accounted for unspecific interactions and focused on DBD-specific differences in DNA recognition. Our results show a stronger AT-dependence for HMGA2, as illustrated in the scatter plots of Figure 7-12b. This normalization is important and will also be used in some of the next sections, for better comparison of HMGA1 and HMGA2 binding.

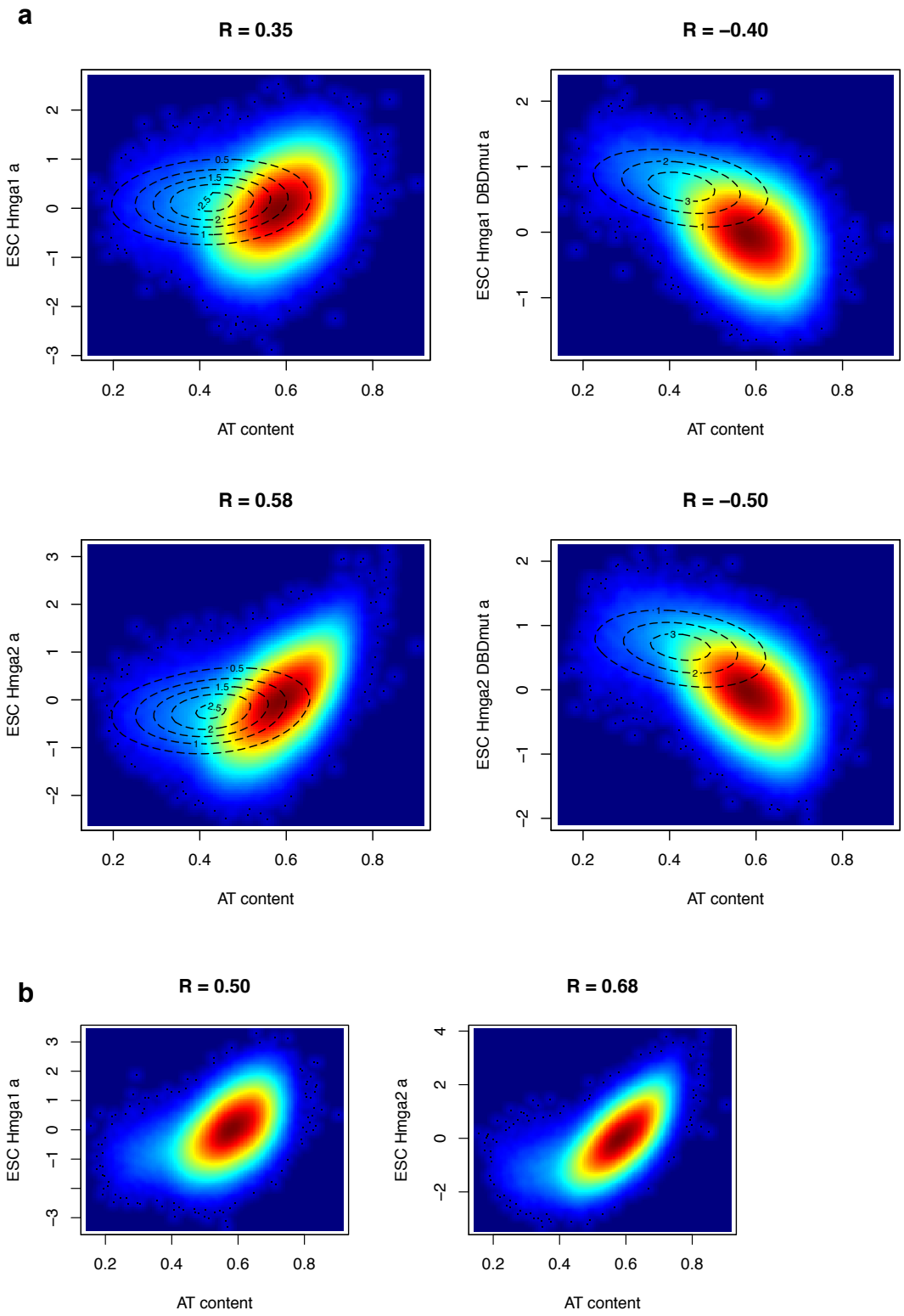


Figure 7-12 Pearson's correlation coefficients and scatter plots of bioHMGA samples vs. AT contents in ES cells. a) Values as log2(IP/Input) over 1kb tiling windows. The contours indicate the density distribution of windows overlapping CGI; b) Scatter plot of AT content vs. HMGA1-2 input-normalized enrichment values over enrichment over input of respective DBDmutant. Same regions as in a)

In the following analysis we investigated potential colocalizations of HMG1-2 with chromatin features other than AT-content. From the results of the PCA one would expect much lower correlations values for any of such features (see Figure 7-11b). We therefore focused only on data generated in our laboratory and in the same model system in order to minimize variability.

By contrasting binding to such chromatin marks and factors, no significant correlation is manifest other than with AT, as summarized by genome-wide correlation plot for both HMGA1 and HMGA2 (Figure 7-13a). For one replicate of HMGA2 the same information is also shown as scatterplots (Figure 7-13b), which illustrate the pattern of individual relationship over the continuum of HMGA2 signal.

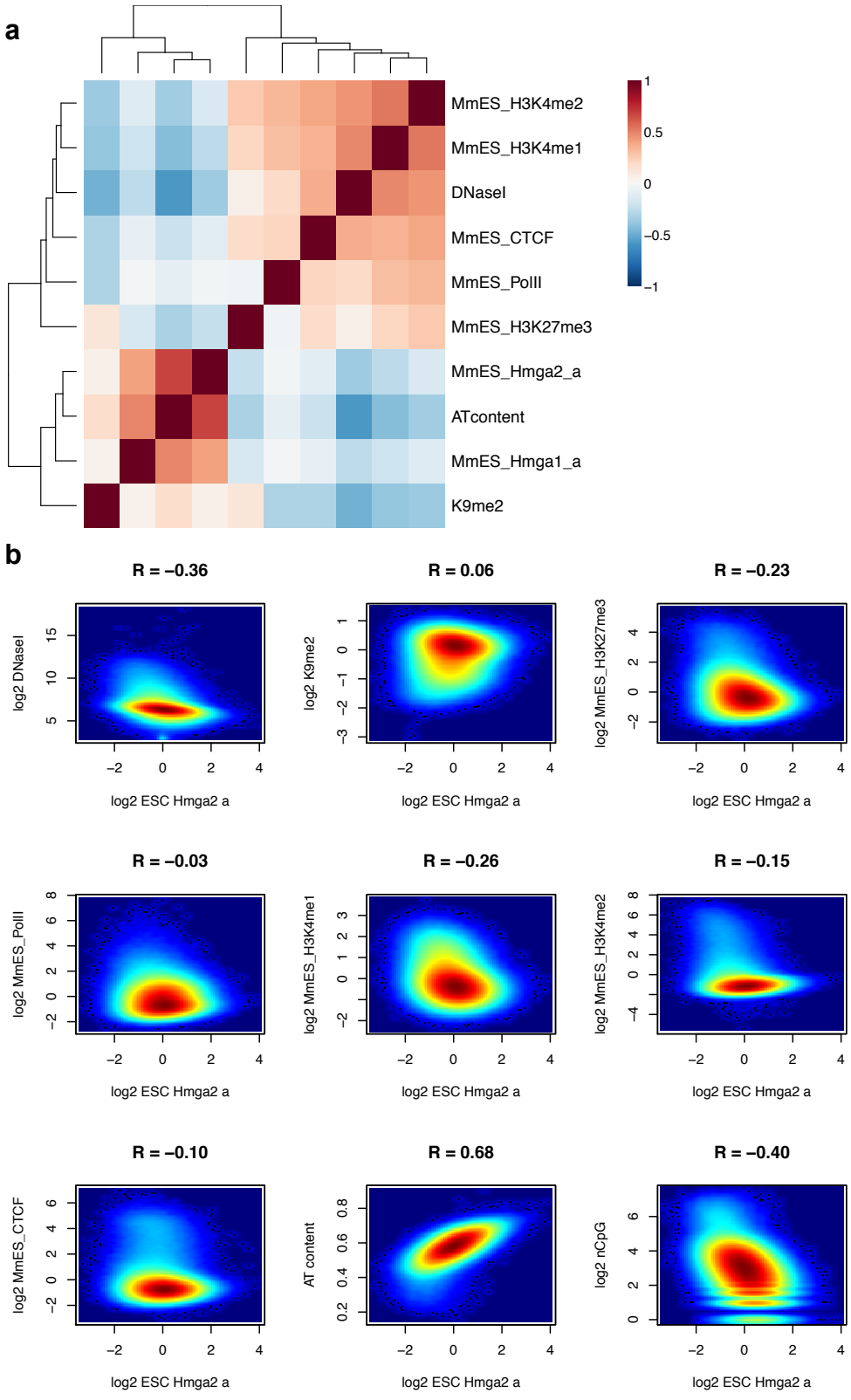


Figure 7-13 a) Unsupervised hierarchical clustering of genome-wide correlations between DNaseI cut frequency, AT content and enrichment over input values for the indicated ChIP or bioChIP samples (1kb tiling windows, colour legend refers to the Pearson correlation coefficient); b) scatter plots of HMGA2 replicate a. Enrichment values against various chromatin associated proteins and genetic features are shown (1kb tiling windows, R: Pearson correlation coefficients).

7.3.4 Binding in different local and global chromatin environments

The results from the previous experiments highlighted that HMGA1-2 proteins bind to DNA in a DBD-dependent manner and that binding correlates genome-wide with AT-richness. We also showed that additional chromatin cues showed significantly lower correlations with HMGA1-2 genome-wide.

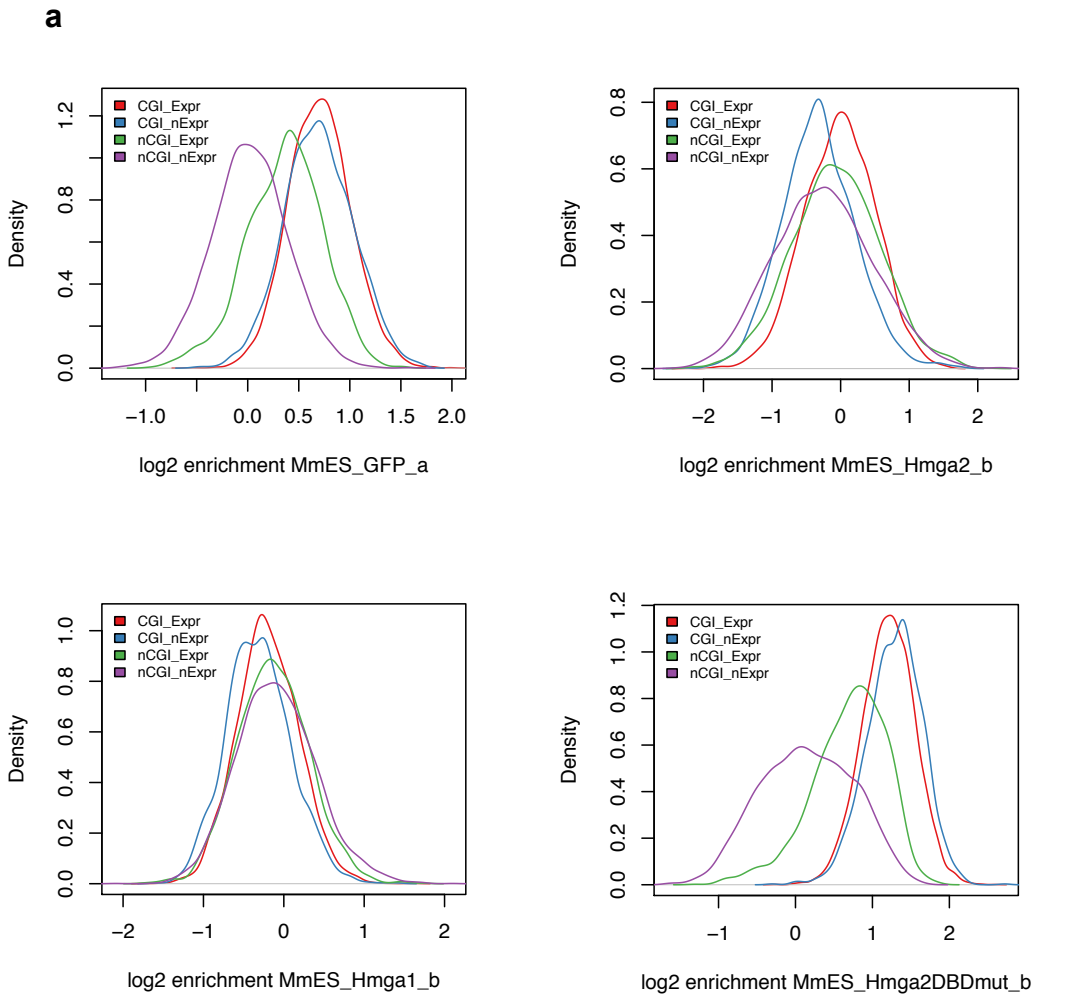
We next asked whether in subsets of genomic regions, a different chromatin environment could have an impact on binding.

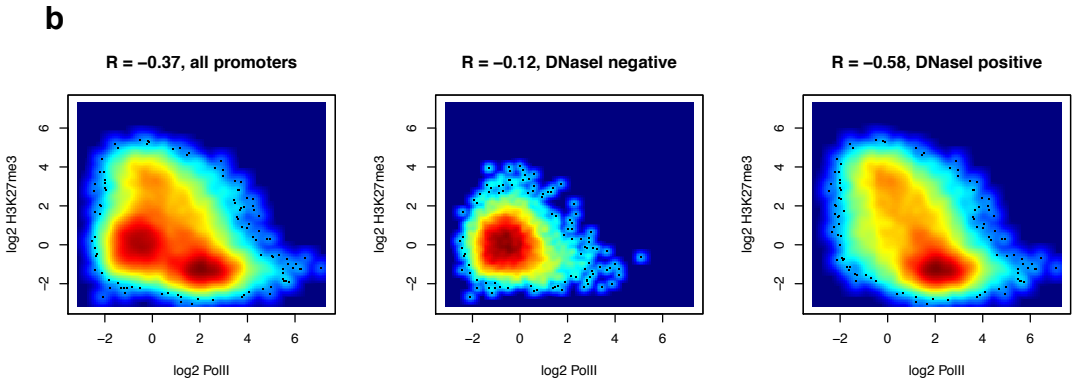
Also, we investigated whether the binding profile was stem-cell specific by repeating our experiments in a differentiated cell type.

7.3.4.1 Impact of chromatin states at differentially transcribed regions

It is generally accepted that the majority of DNA binding factors are sensitive to the chromatin environment of a given locus and those that are not, are referred to as pioneer TF (Beato and Eisfeld, 1997; Wang et al., 2012; Zaret and Carroll, 2011). It is also known that transcriptional status of a given gene correlates with specific chromatin marks (Smolle and Workman, 2013). This notion implies that the chromatin structure at transcribed genes is different from the one present at silent ones.

We therefore asked whether the different chromatin associated with the transcriptional states could affect HMGA1-2 binding. When looking at non CpG island promoters with different activity, DBDmut and GFP are enriched at expressed genes, whereas HMGA1-2 signal is stable (Figure 7-14a). At CpG island promoters, GFP and DBD mutant controls are enriched both at expressed and not expressed genes, in accordance with the observation that accessibility is high at active and Polycomb target promoters, the latter being largely CGIs (panel Figure 7-14b (Jermann et al., 2014; Mendenhall et al., 2010; Schübeler, 2015). Taken together these results are compatible with the explanation that accessible DNA is a preferred point of contact for inert proteins but not for HMGA1-2. In a representative replicate, correlation with DNAseI sensitivity was indeed prominent for HMGA1 DBD mutant but not for HMGA1 at promoters (Figure 7-14c).


a
CGI_Expr
CGI_nExpr
nCGI_Expr
nCGI_nExpr
Density
log2 enrichment MmES_GFP_a
log2 enrichment MmES_Hmga2_b
log2 enrichment MmES_Hmga1_b
log2 enrichment MmES_Hmga2DBDmut_b


b
R = −0.37, all promoters
R = −0.12, DNaseI negative
R = −0.58, DNaseI positive
log2 H3K27me3
log2 PolII

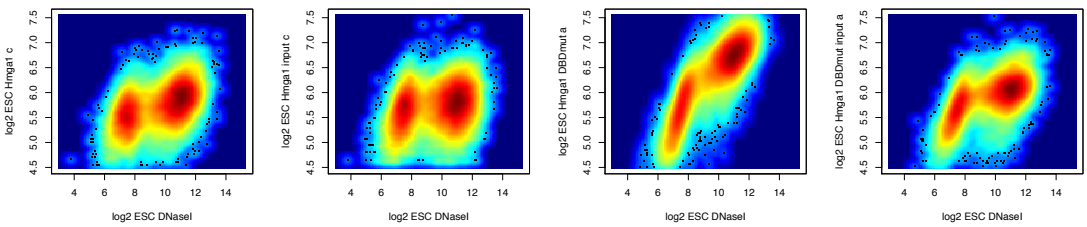
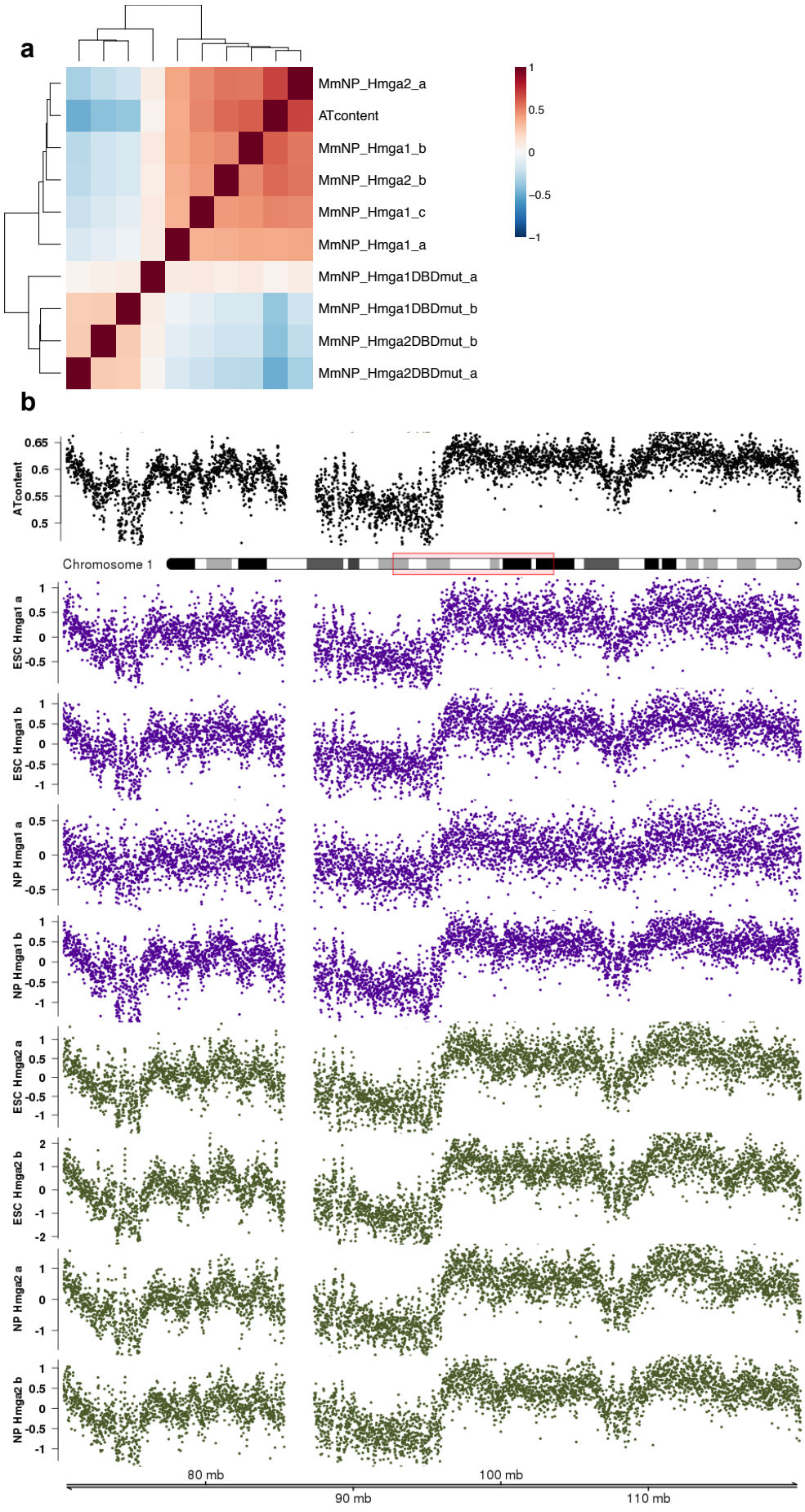

c
log2 ESC Hmga1 c
log2 ESC Hmga1 input c
log2 ESC Hmga1 DBDmut a
log2 ESC Hmga1 DBDmut input a
log2 ESC DNaseI

Figure 7-14 a) Distribution of log2 enrichments (IP/input) for the indicated samples over selected promoters (nCGI/CGI = non or CGI promoters (according to the UCSC CGI annotation), Expr/nExpr = expressed or not expressed (defined via the modes of the bimodal distribution of expression levels)); b) scatterplot of RNAPol2 and H3K72me3 enrichment values at all promoters (left) or at DHS positive or negative promoters (defined analogously as for expression levels); c) Scatterplot of DNaseI cut frequency versus selected IP over input signal at promoters (log2 transformed values).

7.3.4.2 Assessment of genome-wide location in neuronal versus stem cells

All our previous findings were obtained in ESC. However it was important also to assess HMGA1 and HMGA2 genomic location in a different cell type. First, because in committed cells there may be differences in the composition of chromatin, a prominent example being the reduction of H3K4me3-H3K27me3 bivalent domains (Laugesen and Helin, 2014; Mikkelsen et al., 2007). Second, some regions of the genome go through chromatin remodeling during differentiation (Hemberger et al., 2009). Lastly because, at least in the differentiation paradigm that we adopted, endogenous HMGA2 becomes expressed in neuronal progenitor cells (NPC) (see Figure 7-9a). The protocol allows differentiation of ESC towards multipotent Pax6-positive radial glial cells, that can be further differentiated into postmitotic glutamatergic neurons (Bibel et al., 2004). In agreement with above statements, upon differentiation with this system, we also know that ESCs undergo extensive remodeling of histone marks, DNA-methylation and replication timing (Hiratani et al., 2008; Mohn et al., 2008; Stadler et al., 2011). This opened the opportunity to examine HMGA genomic location in a different cell type and in a different chromatin landscape.

We contrasted the binding of HMGA1-2 between ESC and NPC by calculating correlations over genomic windows (**Figure 7-15**a). The high degree of correlation between samples and AT-content indicates superimposable binding at the majority of sites. This supports the observation that primary DNA sequence is effectively the main determinant of genomic location of HMGA proteins. AT-dependence and invariance in binding can also be appreciated visually by looking at the binding pattern along the same stretch of Chr1 as in figure 6-10 for two HMGA ESC and NPC replicates (**Figure 7-15**b). Importantly these findings also suggest that HMGA1 and HMGA2 binding may be conserved in other cell-types, potentially even cancer cells, irrespective of the pathological or physiological epigenetic state.



(legend on next page)

Figure 7-15 a) Correlation plot of NPC enrichment over input for HMGA1 and HMGA2 replicates and their respective DBD mutants. Analysis on 1kb tiling windows. b) Average signal intensity of input normalized log2 enrichment over DBDmutant signal for the indicated samples over the same fraction of Chr1 as in figure 6-10. Each datapoint is calculated for a 10kb tiling window (figure on previous page). For better readability, top and bottom 1% of data range is not shown

7.3.5 Correlation of HMGA proteins with broad and stable chromatin features

After having established that binding of HMGA1-2 does not change with changes in chromatin we asked whether enriched regions coincided with chromatin features that are known to be invariant. Such invariant features are found at constitutive heterochromatin, which is characterized by having low histone acetylation, high H3K9me2 and high cytosine methylation (C David Allis, 2014). Additionally these regions also tend to replicate their DNA late in the cell cycle, even though not all late replicating region represent constitutive heterochromatin (Hiratani et al., 2008). Importantly with respect to our work, constitutive heterochomatic regions show higher than average AT-content due to high prevalence of major and minor satellites repeats and transposon integration events (Lehnertz et al., 2003). In this regard the question we asked is a relevant one, because if HMGA and constitutive heterochromatin co-localize, hints on causality over localization and recruitment can be deduced.

Heterochromatic regions are found at telomeres and centromeres (in mouse all chromosomes are acrocentric) however they can also be found along the chromosomes. Such regions are usually large and are therefore best observed at the chromosomal scale. For this reason, while examining HMGA1-2 enrichments over heterochromatin we looked at 10 kb windows (Figure 7-16a).

After AT-content the highest correlation was observed with LaminA (Figure 7-16b), which locates at the inner nuclear membrane, a well known spatial organizer of heterochromatin (Mattout et al., 2015). The data also showed that the majority of HMGA-high regions are also H3K9me2-high and late replicating.

As an exception, at the locations highlighted by the arrow in Figure 7-16a there is a disconnection between other HMGA1-2 and heterochromatic features. Thus it seems that HMGA1-2 binding and heterochromatin formation and maintenance are probably uncoupled events.

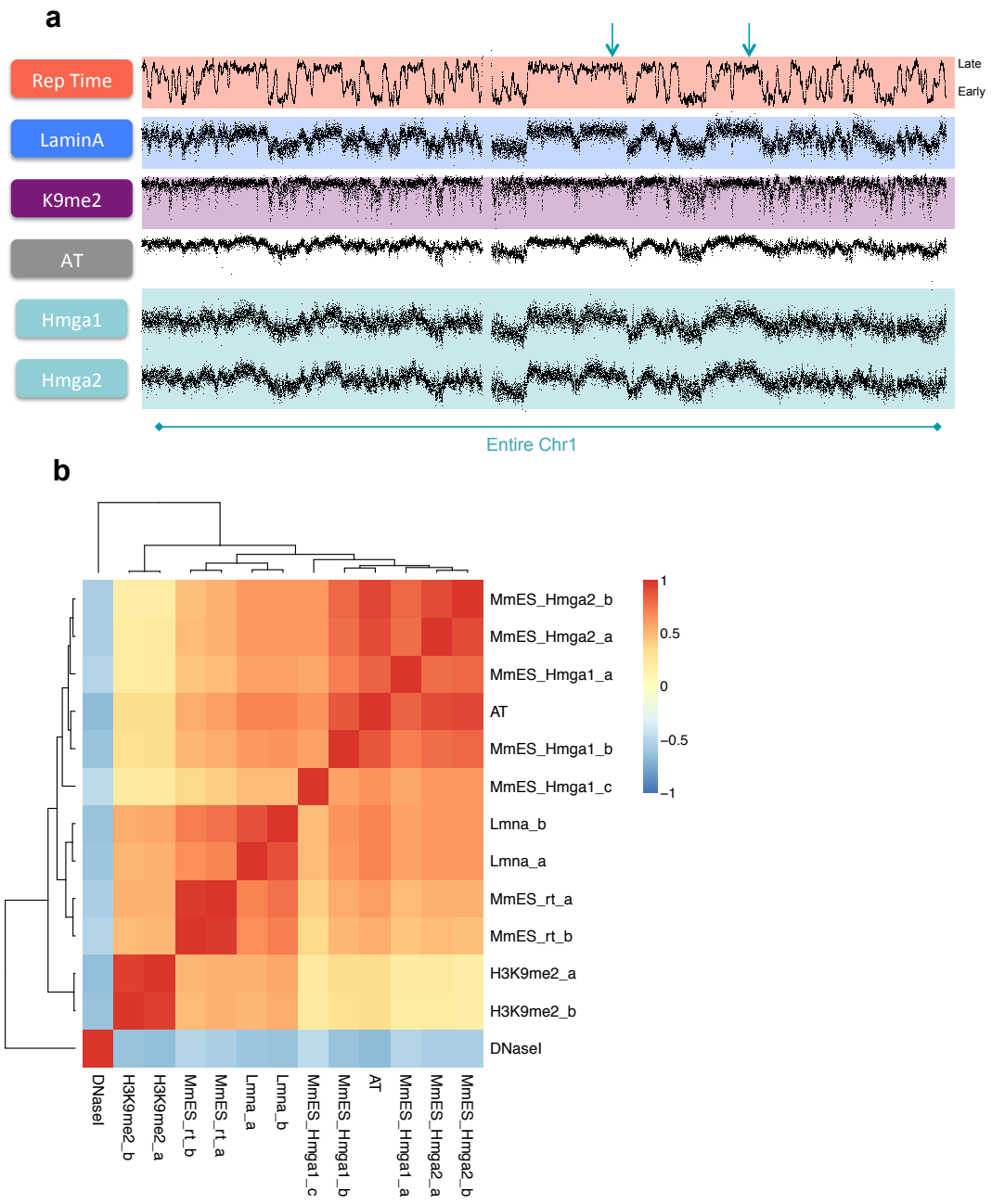


Figure 7-16 Chromosome-wide profiling of the indicated genomic and epigenomic features. Each datapoint represents average signal over 10kb tiling window (Lmna = DamID LaminA sample, HMGA = input normalized enrichment over DBDmutant signal, K9me2 = GSM1314605 H3K9me2 enrichment over GSM1314606 input signal); b) Unsupervised hierarchical clustering and genome-wide correlation of Replication Timing, DNaseI cut frequency, AT-content scores and enrichment values for the indicated samples and (10kb tiling windows, colour legend refers to R, Pearson´s correlation coefficient)

7.3.6 Assessment of Hmga1 KO phenotype and bioChIP experiments in HMGA1-2 add-backs cell lines

Previous work had suggested that HMGA1 might act as a co-activator in vitro and in vivo by stabilizing pre-initiation complex and the enhancesome, the enhancer

associated protein complex contacting active promoters (Reeves et al., 2000; Xu et al., 2011; Yie et al., 1999). Our stem cell lines do not express Hmga2 whereas Hmga1 is expressed to levels comparable to Sox2 (see Figure 7-9). After generating Hmga1 isogenic KO and we could test whether loss of HMGA1 in ESC would cause transcriptional deregulation, as one would expect by KO of an activator.

Our results show that only 18 genes were significantly altered upon Hmga1 KO (Figure 7-17a). Of note, Hmga2 was not one of the upregulated. Accordingly, we could not observe growth defects, morphological alterations or defect in neuronal differentiation (data not shown). Also at the level of transcripts originated from repeats we found no deregulation (Figure 7-17b). Thus we can conclude that HMGA1 has limited role in transcriptional regulation, consistent with its AT-rich/heterochromatin co-localization and its depletion at regulatory regions.

Taking advantage of the Hmga1 KO cell line we went on to test whether bioHMGA binding in the KO background was conserved. In fact before having this control, formally we could not test whether in the WT background endogenous HMGA1 was excluding tagged HMGA1-2 from some sites. Therefore we reintroduced either HMGA1 or HMGA2 proteins in the KO background and repeated bioChIP experiments. Importantly, HMGA1 protein expression was restored to levels comparable to WT as can be observed Figure 7-17c. Subsequent HMGA1 and HMGA2 bioChIP experiments showed superimposable genome-wide distribution (Figure 7-17d) indicating absence of reproducible changes from experiments in the WT background.

These results add consistency to our findings in the WT background and remark the efficacy of RAMBiO approach.

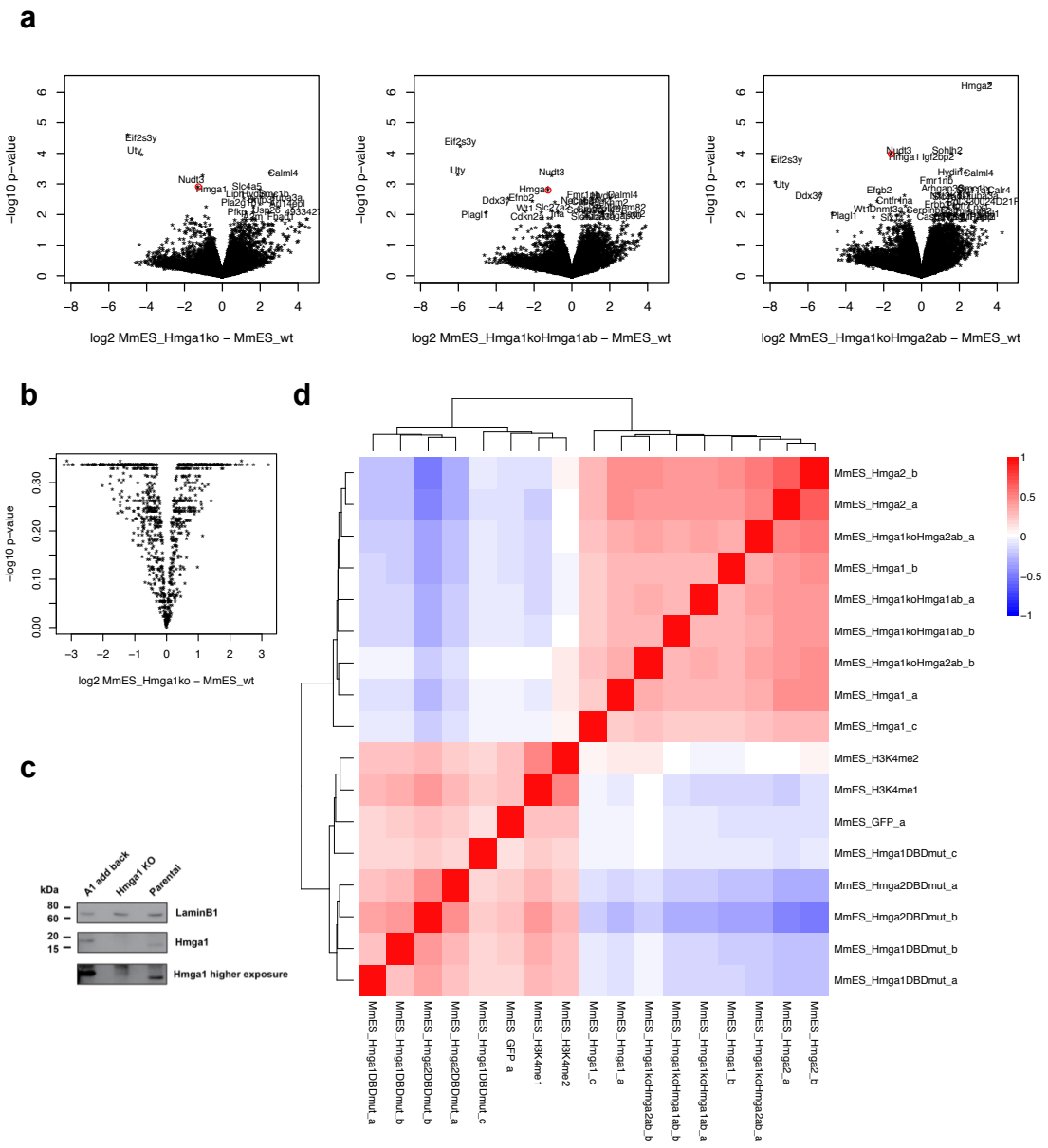


Figure 7-17 a) Volcano plot of the depicted samples vs. WT, gene names are indicated if the gene is significantly differentially expressed (adjusted p-value >= 0.01 and a absolute fold-change of at least 2). Hmga1 is shown in red (analysis min 2 replicates, max alignment repetition = 100, Limma Voom R package). Hmga1 gene, which harbors a frameshift mutation in the Hmga1 KO experiment, remain highly transcribed due to a high level of transcription at untranslated regions; b) Volcano plot for changes in expression at repeats as defined by RepeatMasker http://www.repeatmasker.org/ (log2 transformed values, counting only unique alignments). There are no significant changes (adjusted p-value >= 0.01 and a absolute fold-change of at least 2); c) WB of indicated KO and add-back ES cell lines (30 µgr of total cell protein extracts are loaded per lane); d) Unsupervised hierarchical clustering of genome-wide pairwise correlations of enrichment over input values for the indicated samples (1kb tiling windows, colour legend refers to R: Pearson correlation coefficient). ab = protein add-back in the KO background.

8 Discussion

8.1 Benchmarking RAMBiO performance with a panel of TFs

Using a panel of 12 TF as a proof of principle, we established that RAMBiO can be used to learn the general binding principles of TFs. This approach relies on one-step generation of cell lines expressing TF of interest from defined genomic locations, thus improving state-of-the-art biotin-tagging techniques for location analysis of TF. RAMBiO allows a more straightforward screening of clones expressing the construct of interest as no variegation in gene expression is expected. We show that by screening as few 12 clones per constructs, positive clones are readily detected in two thirds of the cases. This is a remarkable result given the known transcriptional impact of TF.

In the future, simple optimizations in the pipeline could potentially achieve even better throughput. Non-retrieved clones may have encountered growth disadvantage before initiation of the negative selection. An alternative albeit less probable explanation is inefficient biotin conjugation by the BirA enzyme. An improvement that has been already implemented in the host laboratory combines the advantages of RAMBiO with the control of gene expression that can be achieved with inducible promoter systems. By integrating a Tet-On 3G activator and transfecting constructs driven by its responsive element, tight and robust inducible expression is achieved (unpublished data). By using this approach it will be possible to activate TF expression few hours before performing bioChIP experiments thus avoiding confounders caused by sustained TF expression.

Gateway (Invitrogen™) cloning adaptation would be an additional improvement towards conversion of the approach in a high-throughput method for TF location analysis. There is a TF expression library already available in such format (Gubelmann et al., 2013) and the only step needed for RAMBiO application would be introduction of the in frame biotin-tagging peptide.

8.1.1 Observed results for TF binding in mouse ESC

For the majority of the TF expressing clones we obtained high quality bioChIP maps. A comparison with previously generated Sox2 Ab based ChIP-sequencing data showed high similarity of biotin tagging with the most common in vivo TF mapping technique.

We first conducted analysis of the distribution and properties of TF peaks for Sox2, Foxo3 and Smad4. We show that Sox2 binds preferentially to enhancer regions, whereas Foxo3 and Smad4 distribute homogeneously across regulatory regions. This observation seems to agree with a computational analysis of Sox2 distribution based on other datasets. In this publication the authors show that the fraction of bound sites out of in silico predicted is much lower for Sox2 as compared to other TFs (Kuznetsov et al., 2010).

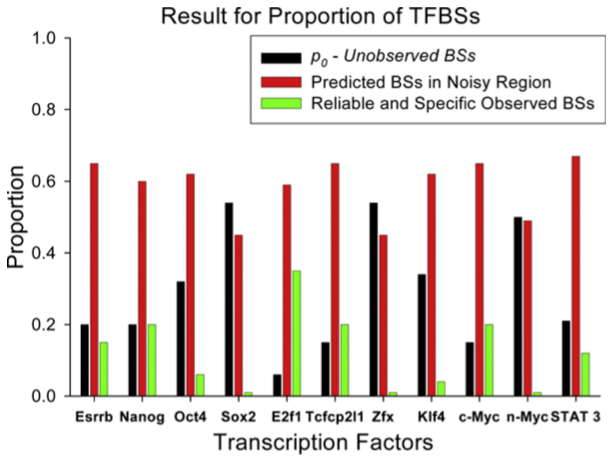


Figure 8-1 Three segments in the range of TF-DNA binding events count for 11 TFs. Data was generated in mouse E14 embryonic cells (GSE 11431). From (Kuznetsov et al., 2010)

However it is very hard to disentangle why a TF like Sox2 in vivo is not binding to its cognate sites at promoters. One explanation could be because of a different chromatin signature, alternatively because of the binding sites syntax (spacing and composition) at that portion of the cistrome. To answer such question, it would take to harness either of the two: the chromatin, with the inherent difficulties in generating histone marks functional KO, or the primary sequence, with having to design a parallel multi-combinatorial protein-DNA interaction study.

Our analysis also highlighted that out of the peak regions identified it was possible to identify TF motifs with good specificity and sensitivity. REST data served as a control due to previously observed high motif enrichments for this Zinc-finger TF, which targets sequence up to 21 bp long (Rockowitz et al., 2014). When we examined the enrichment of Sox2 motifs we saw comparable result for data generated by us and previously available. For Foxo1 and Foxo3, enrichments of a redundant Fox motif were even higher. This result is in agreement with the notion that FOX proteins can be categorized as either pioneer factors (e.g. FOXA subfamily), transcription factors (e.g. FOXP subfamily), or like the FOXO sub-family as having both pioneering and classic transcriptional functions (Lalmansingh et al., 2012). This result may also indicate that the Forkhead DBD has an affinity for its cognate motifs higher than Sox2, however further studies would be needed to confirm the hypothesis. Importantly, such hypothesis can only be advanced since the pull down of chromatin is performed using the same SAV-biotin chemistry.

For Smad proteins we could not see reproducible enrichment of Smad motifs. A possible explanation to this observation is the fact that binding of Smad proteins to DNA, and in particular Smad3, is a highly regulated process downstream of TGF-beta induced signaling and PTM (Heldin et al., 1997).

Altogether our results show that expression and characterization of regulatory region binders is possible and effective and this prompted us to investigate the genomic location of HMG proteins.

8.2 Genomic location analysis of HMGB proteins in the mouse warrants caution when drawing functional conclusions

HMGB proteins have been described as TF that regulate both specific gene transcription and also genome stability by interacting with DNA, histones, other TF and nuclear proteins (Bianchi and Agresti, 2005).

Given the many proposed mechanisms of action we wanted to test the genomic location of the different members of HMGB class of protein. We did so in an unbiased fashion by taking advantage of RAMBiO approach, previously mentioned.

To evaluate functionality of the constructs we assessed subcellular localization of HMGB1 protein and our imaging data show nuclear staining and nucleolar accumulation. This is in agreement with previous imaging on fixed cellular preparations for HMGB1-2-3 (See section 6.2). For HMGB4 we reproduced the previously observed exclusion from DAPI dense foci (Catena et al., 2009). These results indicate that a N-terminally positioned biotin tag is not affecting the correct folding and sub-cellular localization of HMGB1. As HMGB1 contains 2 nuclear localization signals (NLS) and the first of them is embedded in the first HMG-box at aa 27 (Youn and Shin, 2006), a biotin tag at the beginning of the sequence is also probably exposed and does not interfere with HMGB1sub-cellular targeting.

For HMGB1-2-3 the binding profile is highly similar, whereas for HMGB4 we observe binding in the same regions where HMGB1-2-3 accumulate, but the dynamic range in the signal is much higher. Since ChIP-sequencing is a cell population method, this can either mean that the background signal outside bound regions is lower for HMGB4, or that the affinity for target regions is higher. Since a lower amount of DNA was recovered for the HMGB4 sample after ChIP (starting from the same number of cells) we favor the first hypothesis.

In general, HMGB proteins are thus found enriched at open regions in the DNA. However by definition such regions are defined as being more readily contacted by a sequence unspecific enzyme DNAseI. This implies that whenever a protein shows a similar distribution to DHS genome-wide a careful assessment over the specificity of binding should be considered. In this regards, our sequencing data

for a biotin tagged GFP or for a DBD mutant version of HMGB1 (Jung and Lippard, 2003) showed very similar DNA contact profiles to that of WT HMGB1.

In our analysis is also included an antibody ChIP-sequencing experiment for HMGB1, which again shows accumulation at accessible regions. This finding for HMGB1 Ab ChIP well agrees with a publication that appeared in the literature describing HMGB2 binding at active regulatory regions in two different cancer cell lines (Redmond et al., 2014). However given our previous controls, we can be highly skeptical in implicating HMGB1 accumulation at open regions with any kind of regulatory activity. Collaterally, it is interesting to notice the in the initial descriptions of HMGB1-2 properties, authors acknowledge the weak interaction with linear B-DNA (Johns, 1982), with affinities in the order of 5×10^{-5} M (Stros, 2010).

One detail that was highlighted in the Redmond et al. 2014 publication was the fact that in order to get high quality maps the crosslinking protocol had to be changed from 10 min at RT to 1h at 4°C. In light of this observation we show data for a GFP and HMGB2 bioChIP adopting this modified protocol, however we report no difference in terms of improved signal to noise ratio. This result indicates that HMGB2 bioChIP does not benefit from this tweaking in our hands. One possible explanation is that our HMGB2 is not active, due to the difference in cell type or presence of a tag. However an alternative explanation is that the increased signal observed by Ab ChIP is an artifact of the longer crosslinking time, and this is not visible by bioChIP thanks to higher stringency in beads washing.

In the results section a further investigation on the reasons of a wider dynamic range for HMGB4 is presented. We show that after GFP signal subtraction HMGB4 binding is still proportional to DNA accessibility. Accordingly, upon differentiation of ESC towards neurons, HMGB4 signal follows the changes happening at regulatory regions. The motifs that are found below HMGB4 enriched regions are also different, reflecting the advent of an alternative set of TF driving NPC transcription and accessibility.

One obvious difference between HMGB4 and HMGB1-2-3 in terms of nucleotide sequence is the absence of the acidic tail (see Introduction). With the HMGB1 truncation experiment we asked whether such truncated protein would contact

the genome in a manner more similar to HMGB4, however this was clearly not the case. This experiment also excludes a previously proposed mechanism of modulation of HMGB1 affinity for DNA based on an auto-inhibitory interaction between the HMG-boxes and the acidic tail (Lee and Thomas, 2000) (Stott et al., 2014) (Wang et al., 2007). We propose that a possible cause of the observed HMGB4 distribution profile may lie in a different NLS structure (positively charged central basic stretch at aa 80 for HMGB4) or in the aa-sequence differences throughout the HMG-boxes.

Further validation experiments for our bioHMGB1 findings are presented in the last section. Upon CRISPR-KO of Hmgb1 cells were analyzed by transcriptional profiling, however little significant changes are observed. This also contrasts previous models, mainly based on sporadic observations, implicating HMGB1 in controlling gene expression (summarized in this review (Stros, 2010)) and findings in HeLa cells, where a reduction in histones expression was reported upon HMGB1 KD (Celona et al., 2011). We cannot exclude a buffering of the phenotype through HMGB2 (which is also expressed in ESC), but based on our genome-wide binding data it seems improbable.

If the KO experiment were causing a strong transcriptional signature, upon reintroduction of bioHMGB1, we could have had unambiguous proof of functionality in case of phenotypic rescue. Since no transcriptional phenotype was present in the first place, we also performed an additional validation experiment. We compared binding of a C-terminally tagged HMGB1 with the binding of the N-terminal construct discussed so far. The binding profile is very similar, a result that underscores how our bioChIP findings most likely apply to functional proteins. Alternatively both tagging strategies should have caused protein malfunctioning; however this latter scenario does not agree with previous experiments performed on C-terminally YFP-tagged HMGB1 (Agresti et al., 2005).

Collectively our bioChIP experiments indicate that HMGB proteins bind open region of the genome via random DNA contacts. Therefore, reported HMGB2 association with chromatin in vivo and HMGB1 with DNA in vitro

(Pallier et al., 2003; Redmond et al., 2014; Stros, 2010) is not captured by our stringent genomic approach. It is interesting to notice that the strongest affinity

observed for DNA in vitro is with damaged or distorted structures (Bianchi et al., 1989; Ohndorf et al., 1999) that are not normally present in living cells. Perhaps a role in histone chaperoning and shuttling histone protein in proximity to genomic DNA, without direct contacting it, should be reconsidered. It is in fact known that HMGB1 takes only 1-2s to cross the entire diameter of nuclei (Scaffidi et al., 2002), much faster than H1 for example whose residency time on chromatin is 4 min (Stros, 2010). Seminal work and more recent experiments already point in this direction of a direct interaction between HMGB1 and histone proteins (Bonaldi et al., 2003; Zhuang et al., 2014).

The protein HMGB4 seems to contact the same open regions as and its frequency map follows even more closely DNA accessibility.

Although the statements above are supported by numerous observations we acknowledge the difficulty of controlling all possible confounders when describing absence of binding. In the next section, we discuss our findings for HMGA proteins that on the contrary are supported by easily interpretable DBD-mutant experiments, where specific binding is abolished.

8.3 Genomic location analysis of HMGA proteins reveals a unique DNA binding modality

HMGA proteins have been identified due to their property to bind to alfa-satellite sequences in vitro (Strauss and Varshavsky, 1984). These repeat elements show a compositional bias for AT DNA. From this initial observation the DBD was called AT-hook. The authors observed strong DNAseI protection over stretches of 5 or more consecutive A or T nucleotides. However they argued that affinity of the AT-hook for DNA is probably generally high because of extensive interaction with DNA phosphate backbone (Figure 8-2a). Indeed the presence of a conserved positively charged RGR core and a variable number of R or K on either side make the AT-hook a very strongly charged DBD.

These initial findings were replicated when looking at chromosomal banding and colocalization of HMGA1 with repetitive, AT-rich DNA (Disney et al., 1989). In vivo HMGA1-2 localize at DAPI dense foci, which enrich for satellite DNA of telomeres and centromeres.

Following the initial observation many studies replicated HMGA proteins binding to AT-rich DNA in vitro (Reeves and Beckerbauer, 2001). However all evidences were coming from isolated observations or at single loci. In vitro, a low-throughput SELEX assessment of HMGA2 preference returned a high-information-content DNA logo that was never anticipated before (Figure 8-2b, see figure legend).

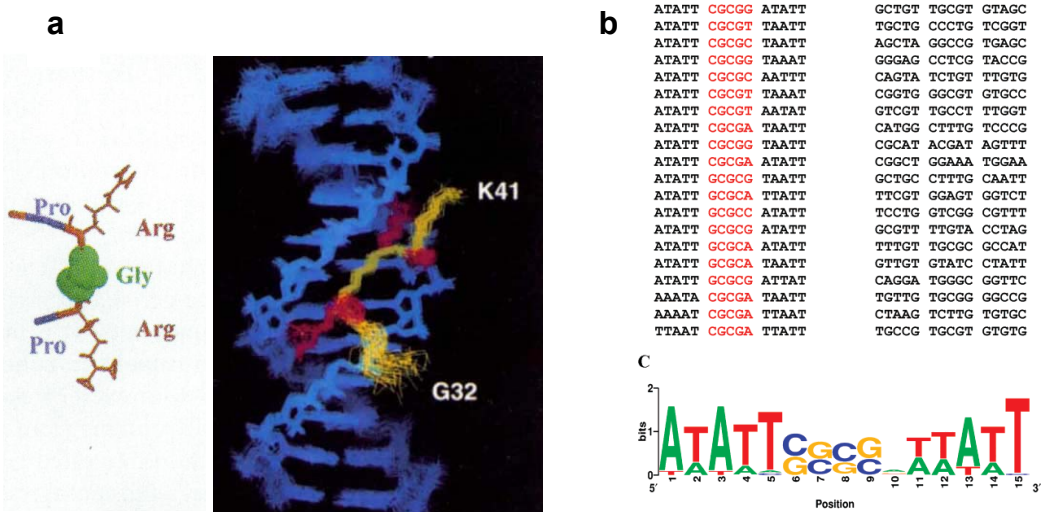


Figure 8-2 a) Left, the core aa composition of the AT-hook DNA binding domain (Reeves, 2000). Right, NMR structure of AT-hook domain in complex with a DNA substrate. In red are Pro or Arg side chains of the conserved aa core (Huth et al., 1997). b) HMGA2 SELEX sequencing result of the top 20 sequences out of 102 sequenced and 71 unique PCR clones after enrichment. On the right, 20 sequences sampled out of the DNA library pre-enrichment(Cui and Leng, 2007).

These in vitro findings could not be replicated in the first study where HMGA2 location was determined in vivo: the identified consensus motif there, is a repetition of W nucleotides (A or T) (Winter et al., 2011) (see section 4.3.3.1. This result however was obtained by sequencing only a few dozens of bacterial clones after plasmid transformation of ChIP extracted DNA and has to be considered a preliminary finding.

In another study a ChIP-chip experiment was performed on MKN28 gastrinoma cell line overexpressing HMGA2 (Zha et al., 2012), but no results were included with respects to in vivo binding preferences (data is not publicly available). More recently, a HMGA2 ChIP-seq study in MEF was published, but again no comments on HMGA2 sequence specificity are made (Singh et al., 2015). Looking at the data, it looks like the antibody that was used gave a promoter-centered enrichment, in clear disagreement with our findings.

In order to shed light on the in vivo binding preferences of HMGA1 and HMGA2 we set out to apply RAMBiO for the study of this controversial class of proteins.

When we expressed HMGA1 and HMGA2 we saw the typical DAPI dense staining for both proteins (Disney et al., 1989; Harrer, 2004; Henriksen et al., 2010). We were pleased by the result as we were aware of reported wrong localization of a N-terminal GFP-HMGA1 fusion construct (Catez and Hock, 2010).

At a first glance on a genome browser, genomic localization of HMGA1-2 is very similar to input DNA. However it is known that a sequencing bias, mainly caused by PCR amplification, is present in modern sequencing results (Figure 8-3). Therefore we reasoned that before drawing conclusions we had to correct for this sequencing bias. By doing so it was possible to appreciate enrichment of HMGA1-2 at specific genomic regions, and indeed the read-count GC distribution for our IP samples was resistant to the drop observed at low GC-content regions in the input fraction. This suggested that HMGA proteins were binding in vivo to AT rich DNA substrates.

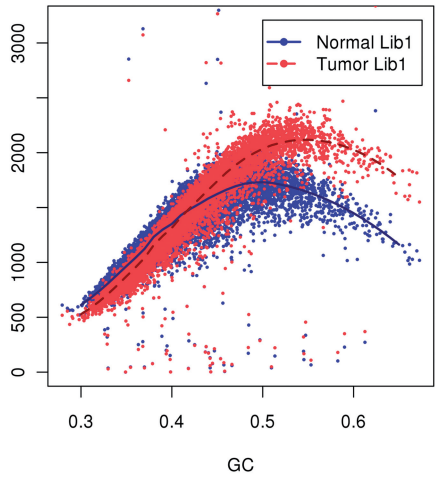


Figure 8-3 Read counts over GC content for genomic human DNA samples. Mean values on windows of 10 kb from Chr1. From (Benjamini and Speed, 2012).

To evaluate the origin of this preference we assessed the impact of mutations in key aa, known to be contacting DNA from in vitro evidence. Mutation of the DBDs of HMGA1-2 caused a reduction in affinity for DNA and loss of the AT rich binding. Therefore we could show that binding is dependent on functional AT-hooks. This is an important result that shows that HMGA1-2 genome-wide distribution is likely not determined by indirect recruitment to chromatin. Collaterally, since mutating the DBD causes a clear change in HMGA1-2 binding, we speculate that HMGA1-2 are probably not affected by biotinylation (also in the light of imaging results and the high pull-down efficiencies).

To evaluate binding determinants in an unbiased approach we subjected the pool of datasets (HMGA proteins, the mutated counterparts and GFP samples) to PCA. We observe that one feature alone can explain a big portion of the variance. In other words a single hidden variable is able to discriminate HMGA proteins from DBD mutated proteins. Since AT-hooks domains bind AT-rich DNA in vitro (Fonfría-Subirós et al., 2012; Singh et al., 2006), we made an informed guess and contrasted the identified PC1 to AT-content to discover a clear correlation.

The observation was confirmed by directly evaluating the correlation between AT content and HMGA1-2 enrichments over input. Both proteins were not enriched (nor depleted) at sequences containing just few A or T (in the range of GC richness of CpG island). The majority of the genome however contains sufficient A or T nucleotides to elicit a linear response at the level of HMGA1-2

binding. This is an interesting observation that is reminiscent of the linear dependence to methylation density uncovered for Mbd proteins, other binders of low complexity DNA motifs (Baubec et al., 2013).

Next we show absence of genome-wide correlation with other chromatin marks and components. This is already suggested by the PCA, with the little contribution of additional PCs in explaining potential differences between HMGAs (mutated or not) and the inert GFP. It follows that HMGA1-2 genomic location is almost entirely encoded in the respective DBDs. This notion prompted us to explicitly compare HMGA1 and HMGA2 DBDs in terms of sensitivity to AT-content. To account for unspecific protein dependent confounders we add an additional normalization step, by subtracting DBD mutant signal from the IP enrichments over input. Our results show very similar AT -regression slope for both proteins and comparable maximal enrichments (4-8 fold in both cases). However for HMGA1 there is a higher noise in AT readout, pointing to either lower affinity or higher sensitivity to chromatin cues.

In this context we asked whether a different chromatin environment was able to modulate affinity to AT-rich DNA at specific subsets of regions. When we looked at differentially transcribed genes we did not see any difference. A similar conclusion was drawn when we looked for modulation of binding in response to the changes in chromatin states that happen during neuronal differentiation.

Altogether it appears that AT-content is the sole determinant of HMGA1-2 binding genome-wide, and not even at subsets of regions binding is affected by chromatin cues like accessibility.

8.3.1 Proportion of A or T nucleotides determines HMGA1-2 binding

Our results have important implications because they reveal the nature of the proposed HMGA1-2 preference for AT-rich DNA. Our data shows that binding occurs throughout the genome over a continuum of affinities, except at CGI where A or T bases are too sparse. It is important to point out that from a biochemical perspective the pattern of hydrogen bond donors and acceptors in the minor groove does not allow discriminating A:T from T:A or G:C from C:G base pairs (Figure 8-4). Thus minor groove binders like HMGA proteins may simply recognize degenerate sequences of the type W^n or W-rich (where W

stands for A or T nucleotides from the IUPAC nomenclature, Weak cross-strand binding interactions).

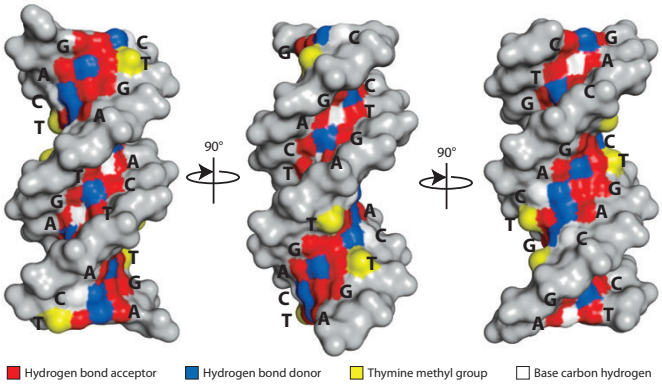


Figure 8-4 Sequence-specific patterns on the edges of the bases in the major groove underlie the ability of proteins to readout base pairs through hydrogen bonds and hydrophobic contacts (hydrogen bond acceptors in red, donors in blue, thymine methyl group in yellow, and base carbon hydrogens in white). In contrast, A:T versus T:A and C:G versus G:C are indistinguishable in the minor groove. From (Rohs et al., 2010).

However AT-rich sequences are also often associated with narrow minor grooves. In particular A-tract, ApT and ApA (TpT) sequences induce narrowing of the minor groove (Rohs et al., 2009). In such cases, arginine mediated recognition of the enhanced negative electrostatic potential offers a mechanism for sequence-specific readout from DNA shape.

At the regions where HMGA1-2 enrichments are highest, the concentration of A and T is so high that it is very difficult to distinguish which of the two mechanisms is preponderant.

What we can conclude is that, since each bioChIP experiment pulls down significant amounts of DNA and because enrichments grow linearly with AT-content, affinity for DNA is high everywhere and is maximal at large AT-rich regions. What is in vivo the minimal motif recognized is a question that we cannot address with standard bioChIP resolution. Potentially, a ChIP-exo adaptation for RAMBiO may better suited to address the question by probing with higher resolution HMGA binding at AT-poor regions.

8.3.2 Transcriptional impact of HMGA1 binding

HMGA1-2 misregulation has been implicated in tumorigenesis (Cleynen and Van de Ven, 2008; Fusco and Fedele, 2007). In malignant epithelial tumors as well as

in leukemia, expression of HMGA1 is upregulated (Morishita et al., 2013). HMGA2 overexpression also causes transformation (Sun et al., 2013). Hmga genes are often involved in rearrangements, mostly in benign tumors of mesenchymal origin (Henriksen et al., 2010). Often, the proposed mechanism of action involves transcriptional deregulation. However a direct role in transcriptional control has never been exhaustively proven for HMGA genome wide. Locus specific examples have been reported for the activation of IFN-beta and IL-2Ralfa genes (Reeves et al., 2000; Yie et al., 1999). Stabilization of enhancer associated protein complexes and displacement of positioned nucleosomes have been advocated for the activation of IFN-beta and IL-2Ralfa genes. Competition with histone H1 and direct interaction with mediator and histone chaperones are some of the other proposed mechanisms of action (Arnoldo et al., 2015; Reeves and Beckerbauer, 2001; Xu et al., 2011).

In our cellular system, Hmga1 expression was high in ESC and upon introduction of bioHMGA1 or bioHMGA2 we did not observe a growth phenotype.

We generated CRISPR KO cell lines for Hmga1 harboring deletions in the third exon. We profiled transcription by RNA-seq and we could show significant changes for only a minority of genes and no change for any repeat class in ESC. This result is in line with the genomic location that we describe. Indeed in the cells that we investigated, HMGA proteins do not enrich for regulatory regions as recently reported in a different model system (Singh et al., 2015). We asked where in the genome were the regions of higher AT-content, and thus higher HMGA1-2 binding. We discovered extensive overlap with heterochromatic DNA. In more detail, we noticed for regions of medium-low HMGA1-2 enrichment facultative association with heterochromatin. However regions of higher HMGA1-2 enrichment were almost constantly associated with LaminA, late replicating and high in H3K9me2 histone mark. In general heterochromatic regions tend to be gene poor, and therefore we show that HMGA1-2 are not enriched over regions of high gene density, including regulatory regions.

Also in terms of reconciling the observed association with the proliferative state of cells in homeostasis and disease, we tend to disfavor models that give HMGA proteins a central role in transcriptional regulation (Sgarra et al., 2010). Our findings better align to a recent observation that connects human HMGA1 with

genome organization via proper positioning of chromosome domains (Shachar et al., 2015). Indeed the transcriptional changes observed in cancer cells upon HMGA up-regulation may as well be secondary to mislocalization of genes under topological control of gene expression (Harr et al., 2016).

In the Hmga1 KO cell line we reintroduced either HMGA1 or HMGA2. Even though not significant, upon HMGA1 add-back a trend towards reduction of the deregulation can be inferred. HMGA2 add-back on the contrary increased variability. These results suggest that the little changes that occur upon Hmga1 KO get fixed with time and cannot be reverted by a bioHMGA rescue.

Nonetheless in the HMGA1 add-back cell line we could prove similar binding of HMGA1 as in the WT background. This result speaks against a competition between endogenous HMGA1 and bioHMGA1, being the cause of a higher noise in AT-content read-out for HMGA1 protein.

9 Conclusion and outlook

High mobility group (HMG) proteins have been extensively studied after their initial discovery as the most abundant non-histone chromatin components. Locus specific examples and in vitro work suggested a role in almost every aspect of nuclear function, ranging from gene regulation, to DNA damage and nucleosome remodeling.

In order to correlate function with binding preferences in vivo, we investigated the genomic location of these proteins with an antibody-independent approach. Robust location data for biotin tagged HMGB1-2-3-4 and HMGA1-2 was generated in both mouse embryonic stem cells and neuronal progenitors cells.

We show that the proteins belonging to HMGB family are enriched “open” region in genome, however their binding is not much dissimilar to that of an inert exogenous protein (GFP). Additionally, we demonstrate for HMGB1 that this binding profile is conserved upon mutation of the DNA-binding domain (DBD).

Although we cannot exclude that the biotin tag is impairing HMGB function we are inclined to conclude that HMGB proteins bind DNA weakly or not at all.

Even though further studies are needed, our preliminary observations for HMGB proteins are in agreement with the notion that HMGB1 is one of the fastest diffusing proteins in the nucleus and that it does not reside on DNA for long periods of time.

For HMGA1 and HMGA2 we show strong binding DNA throughout the genome, with a preference for DNA with high A or T content. Since binding to DNA is widespread, to uncover this preference it was crucial to contrast enriched chromatin with sonicated DNA. By applying this contrast we show that A/T-dependence is not affected by chromatin state and HMGA1-2 binding is DBD dependent. Binding profiles are also maintained upon neuronal differentiation, underscoring the robustness of the primary sequence readout.

Accordingly regions that show a compositional bias towards A/T, like heterochromatin (major and minor satellites, simple repeats) show higher occupancy of both HMGA proteins.

Our genome wide results are compatible with previous in vitro observations, imaging data and preliminary ChIP studies indicating preference for A/T rich

DNA. However we show that binding is not happening only at A/T rich sequences but is widespread all over the genome.

To gain further insight into HMGA function we also generated knock out ESC for HMGA1. What we saw was limited transcriptional deregulation. Together with the location maps we disfavor a transcription-centric view over the main function of HMGA proteins.

It will be interesting to assess in future studies whether such a broad binder of the genome, like histones, show enrichment of specific modifications at discrete sites. Also it will be interesting to assess whether cells lacking HMGA proteins show defects in differentiation efficiencies, which theoretically could reconcile the discrepancies in the transcriptional effects that we describe in stem cells versus those reported by others in committed cell types.

Altogether we think that the work presented in this thesis improved our understanding of the molecular properties of HMG A and B proteins, which will enable more rationale design for future experiments aimed at better elucidating their functions in stem and differentiated cells.

10 Bibliography

Adams, C.C., and Workman, J.L. (1995). Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. Molecular and Cellular Biology *15*, 1405–1421.

Agresti, A., and Bianchi, M.E. (2003). HMGB proteins and gene expression. Current Opinion in Genetics & Development *13*, 170–178.

Agresti, A., Scaffidi, P., Riva, A., Caiolfa, V.R., and Bianchi, M.E. (2005). GR and HMGB1 interact only within chromatin and influence each other's residence time. Molecular Cell *18*, 109–121.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Aravind, L., and Landsman, D. (1998). AT-hook motifs identified in a wide variety of DNA-binding proteins. Nucleic Acids Research *26*, 4413–4421.

Arnoldo, L., Sgarra, R., Chiefari, E., Iiritano, S., Arcidiacono, B., Pegoraro, S., Pellarin, I., Brunetti, A., and Manfioletti, G. (2015). A novel mechanism of post-translational modulation of HMGA functions by the histone chaperone nucleophosmin. Sci. Rep. *5*, 8552–8559.

Bancaud, A.E.L., Huet, S.E.B., Daigle, N., Mozziconacci, J., Beaudouin, J.E.L., and Ellenberg, J. (2009). Molecular crowding affects diffusion and binding of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin. The EMBO Journal *28*, 3785–3798.

Bannister, A.J., Zegerman, P., Partridge, J.F., Miska, E.A., Thomas, J.O., Allshire, R.C., and Kouzarides, T. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature *410*, 120–124.

Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S., and Natoli, G. (2014). Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers. Molecular Cell 1–14.

Baubec, T., Colombo, D.F., Wirbelauer, C., Schmidt, J., Burger, L., Krebs, A.R., Akalin, A., and Schübeler, D. (2015). Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature 1–17.

Baubec, T., Ivanek, R., Lienert, F., and Schübeler, D. (2013). Methylation-dependent and -independent genomic targeting principles of the MBD protein family. Cell *153*, 480–492.

Beato, M., and Eisfeld, K. (1997). Transcription factor access to chromatin. Nucleic Acids Research.

Begum, N., Pash, J.M., and Bhorjee, J.S. (1990). Expression and synthesis of high

mobility group chromosomal proteins in different rat skeletal cell lines during myogenesis. J. Biol. Chem. *265*, 11936–11941.

Benecke, A., Eilebrecht, S., Benecke, A., and Eilebrecht, S. (2015). RNA-Mediated Regulation of HMGA1 Function. Biomolecules *5*, 943–957.

Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research *40*, e72–e72.

Bernstein, B.E., Humphrey, E.L., Erlich, R.L., Schneider, R., Bouman, P., Liu, J.S., Kouzarides, T., and Schreiber, S.L. (2002). Methylation of histone H3 Lys 4 in coding regions of active genes. Proceedings of the National Academy of Sciences *99*, 8695–8700.

Bianchi, M.E., Beltrame, M., and Paonessa, G. (1989). Specific recognition of cruciform DNA by nuclear protein HMG1. *243*, 1056–1059.

Bianchi, M.E., and Agresti, A. (2005). HMG proteins: dynamic players in gene regulation and differentiation. Current Opinion in Genetics & Development *15*, 496–506.

Bianchi, M.E., and Manfredi, A. (2004). Chromatin and cell death. Biochim. Biophys. Acta *1677*, 181–186.

Bibel, M., Richter, J., Schrenk, K., Tucker, K.L., Staiger, V., Korte, M., Goetz, M., and Barde, Y.-A. (2004). Differentiation of mouse embryonic stem cells into a defined neuronal lineage. Nature Neuroscience *7*, 1003–1009.

Bickmore, W.A., and van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. Cell *152*, 1270–1284.

Biggin, M.D. (2011). Animal Transcription Networks as Highly Connected, Quantitative Continua. Developmental Cell *21*, 611–626.

Bonaldi, T., Längst, G., Strohner, R., Becker, P.B., and Bianchi, M.E. (2002). The DNA chaperone HMGB1 facilitates ACF/CHRAC-dependent nucleosome sliding. The EMBO Journal *21*, 6865–6873.

Bonaldi, T., Talamo, F., Scaffidi, P., Ferrera, D., Porto, A., Bachi, A., Rubartelli, A., Agresti, A., and Bianchi, M.E. (2003). Monocytic cells hyperacetylate chromatin protein HMGB1 to redirect it towards secretion. The EMBO Journal *22*, 5551–5560.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., et al. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349–353.

Brinkman, E.K., Chen, T., Amendola, M., and van Steensel, B. (2014). Easy quantitative assessment of genome editing by sequence trace decomposition. Nucleic Acids Research *42*, e168–e168.

Burgess, R.J., and Zhang, Z. (2013). Histone chaperones in nucleosome assembly and human disease. Nat Struct Mol Biol *20*, 14–22.

Bustin, M. (1999). Regulation of DNA-dependent activities by the functional motifs of the high-mobility-group chromosomal proteins. Molecular and Cellular Biology *19*, 5237–5246.

Bustin, M. (2001). Revised nomenclature for high mobility group (HMG) chromosomal proteins. Trends in Biochemical Sciences *26*, 152–153.

Bustin, M., Lehn, D.A., and Landsman, D. (1990). Structural features of the HMG chromosomal proteins and their genes. Biochim. Biophys. Acta *1049*, 231–243.

C David Allis, M.-L.C.T.J.D.R.A.M.L. (2014). Epigenetics, Second Edition. 1–14.

Carroll, S.B. (2008). Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell *134*, 25–36.

Catena, R., Escoffier, E., Caron, C., Khochbin, S., Martianov, I., and Davidson, I. (2009). HMGB4, a Novel Member of the HMGB Family, Is Preferentially Expressed in the Mouse Testis and Localizes to the Basal Pole of Elongating Spermatids. Biology of Reproduction *80*, 358–366.

Catez, F., and Hock, R. (2010). Biochimica et Biophysica Acta. BBA - Gene Regulatory Mechanisms *1799*, 15–27.

Catez, F., Brown, D.T., Misteli, T., and Bustin, M. (2002). Competition between histone H1 and HMGN proteins for chromatin binding sites. EMBO Rep *3*, 760–766.

Celona, B., Weiner, A., Di Felice, F., Mancuso, F.M., Cesarini, E., Rossi, R.L., Gregory, L., Baban, D., Rossetti, G., Grianti, P., et al. (2011). Substantial Histone Reduction Modulates Genomewide Nucleosomal Occupancy and Global Transcriptional Output. PLoS Biol *9*, e1001086.

Ciabrelli, F., and Cavalli, G. (2015). Chromatin-Driven Behavior of Topologically Associating Domains. Journal of Molecular Biology *427*, 608–625.

Cleynen, I., and Van de Ven, W.J.M. (2008). The HMGA proteins: a myriad of functions (Review). Int. J. Oncol. *32*, 289–305.

Cuddapah, S., Schones, D.E., Cui, K., Roh, T.Y., Barski, A., Wei, G., Rochman, M., Bustin, M., and Zhao, K. (2011). Genomic Profiling of HMGN1 Reveals an Association with Chromatin at Regulatory Regions. Molecular and Cellular Biology *31*, 700–709.

Cuellar-Partida, G., Buske, F.A., McLeay, R.C., Whitington, T., Noble, W.S., and Bailey, T.L. (2011). Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics *28*, 56–62.

Cui, T., and Leng, F. (2007). Specific Recognition of AT-Rich DNA Sequences by

the Mammalian High Mobility Group Protein AT-hook 2: A SELEX Study †. Biochemistry *46*, 13059–13066.

Čabart, P., Kalousek, I., Jandová, D., and Hrkal, Z. (1995). Differential expression of nuclear HMG1, HMG2 proteins and H10 histone in various blood cells. Cell Biochem. Funct. *13*, 125–133.

Das, D., Peterson, R.C., and Scovell, W.M. (2004). High mobility group B proteins facilitate strong estrogen receptor binding to classical and half-site estrogen response elements and relax binding selectivity. Mol. Endocrinol. *18*, 2616–2632.

Davey, C.A., Sargent, D.F., Luger, K., Maeder, A.W., and Richmond, T.J. (2002). Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution. Journal of Molecular Biology *319*, 1097–1113.

de Boer, E., Rodriguez, P., Bonte, E., Krijgsveld, J., Katsantoni, E., Heck, A., Grosveld, F., and Strouboulis, J. (2003). Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice. Proc Natl Acad Sci U S A *100*, 7480–7485.

de Mendoza, A., Sebé-Pedrós, A., Šestak, M.S., Matejcic, M., Torruella, G., Domazet-Loso, T., and Ruiz-Trillo, I. (2013). Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proceedings of the National Academy of Sciences *110*, E4858–E4866.

Deng, T., Zhu, Z.I., Zhang, S., Leng, F., Cherukuri, S., Hansen, L., Mariño-Ramírez, L., Meshorer, E., Landsman, D., and Bustin, M. (2013). HMGN1 modulates nucleosome occupancy and DNase I hypersensitivity at the CpG island promoters of embryonic stem cells. Molecular and Cellular Biology *33*, 3377–3389.

Disney, J.E., Johnson, K.R., Magnuson, N.S., Sylvester, S.R., and Reeves, R. (1989). High-mobility group protein HMG-I localizes to G/Q- and C-bands of human and mouse chromosomes. The Journal of Cell Biology *109*, 1975–1982.

Domcke, S., Bardet, A.F., Ginno, P.A., Hartl, D., Burger, L., and Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. Nature *528*, 575–579.

Ellwood, K.B., Yen, Y.M., Johnson, R.C., and Carey, M. (2000). Mechanism for specificity by HMG-1 in enhanceosome assembly. Molecular and Cellular Biology *20*, 4359–4370.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Esposito, F., De Martino, M., D'Angelo, D., Mussnich, P., Raverot, G., Jaffrain-Rea, M.-L., Fraggetta, F., Trouillas, J., and Fusco, A. (2015). HMGA1-pseudogene expression is induced in human pituitary tumors. Cc *14*, 1471–1475.

Fedele, M., Pierantoni, G.M., Berlingieri, M.T., Battista, S., Baldassarre, G., Munshi,

N., Dentice, M., Thanos, D., Santoro, M., Viglietto, G., et al. (2001). Overexpression of Proteins HMGA1 Induces Cell Cycle Deregulation and Apoptosis in Normal Rat Thyroid Cells. Cancer Research *61*, 4583–4590.

Field, Y., Sharon, E., and Segal, E. (2011). How transcription factors identify regulatory sites in genomic sequence. Subcell. Biochem. *52*, 193–204.

Filipczyk, A., Marr, C., Hastreiter, S., Feigelman, J., Schwarzfischer, M., Hoppe, P.S., Loeffler, D., Kokkaliaris, K.D., Endele, M., Schauberger, B., et al. (2015). Network plasticity of pluripotency transcription factors in embryonic stem cells. Nat Cell Biol *17*, 1235–1246.

Fonfría-Subirós, E., Acosta-Reyes, F., Saperas, N., Pous, J., Subirana, J.A., and Campos, J.L. (2012). Crystal Structure of a Complex of DNA with One AT-Hook of HMGA1. PLoS ONE *7*, e37120–e37125.

Frank, O., Schwanbeck, R., and Wisniewski, J.R. (1998). Protein Footprinting Reveals Specific Binding Modes of a High Mobility Group Protein I to DNAs of Different Conformation. J. Biol. Chem. *273*, 20015–20020.

Fusco, A., and Fedele, M. (2007). Roles of HMGA proteins in cancer. Nat Rev Cancer *7*, 899–910.

Gaidatzis, D., Lerch, A., Hahne, F., and Stadler, M.B. (2015). QuasR: quantification and annotation of short reads in R. Bioinformatics *31*, 1130–1132.

Gilbert, N., and Allan, J. (2014). Supercoiling in DNA and chromatin. Current Opinion in Genetics & Development *25*, 15–21.

Giraud, G., Stadhouders, R., Conidi, A., Dekkers, D.H.W., Huylebroeck, D., Demmers, J.A.A., Soler, E., and Grosveld, F.G. (2014). NLS-tagging: an alternative strategy to tag nuclear proteins. Nucleic Acids Research *42*, e163–e163.

Gonçalves, I., Duret, L., and Mouchiroud, D. (2000). Nature and structure of human genes that generate retropseudogenes. Genome Research *10*, 672–678.

Goodwin, G.H., and Johns, E.W. (1973). Isolation and characterisation of two calf-thymus chromatin non-histone proteins with high contents of acidic and basic amino acids. Eur J Biochem *40*, 215–219.

Gubelmann, C., Waszak, S.M., Isakova, A., Holcombe, W., Hens, K., Iagovitina, A., Feuz, J.-D., Raghav, S.K., Simicevic, J., and Deplancke, B. (2013). A yeast one-hybrid and microfluidics-based pipeline to map mammalian gene regulatory networks. Mol Syst Biol *9*, 1–18.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910.

Harr, J.C., Sandoval, A.G., and Gasser, S.M. (2016). Histones and histone modifications in perinuclear chromatin anchoring: from yeast to man. EMBO Rep

17, 139–155.

Harrer, M. (2004). Dynamic interaction of HMGA1a proteins with chromatin. Journal of Cell Science *117*, 3459–3471.

He, X., Chatterjee, R., John, S., Bravo, H., Sathyanarayana, B.K., Biddie, S.C., FitzGerald, P.C., Stamatoyannopoulos, J.A., Hager, G.L., and Vinson, C. (2013). Contribution of nucleosome binding preferences and co-occurring DNA sequences to transcription factor binding. BMC Genomics *14*, 428.

Heidari, N., PHANSTIEL, D.H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., Zhang, M.Q., and Snyder, M.P. (2014). Genome-wide map of regulatory interactions in the human genome. Genome Research *24*, 1905–1917.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics *39*, 311–318.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell *38*, 576–589.

Heldin, C.H., Miyazono, K., and Dijke, ten, P. (1997). TGF-beta signalling from cell membrane to nucleus through SMAD proteins. Nature *390*, 465–471.

Hemberger, M., Dean, W., and Reik, W. (2009). Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. Nat. Rev. Mol. Cell Biol. *10*, 526–537.

Henikoff, S. (2008). Nucleosome destabilization in the epigenetic regulation of gene expression. Nature Reviews Genetics *9*, 15–26.

Henriksen, J., Stabell, M., Meza-Zepeda, L.A., Lauvrak, S.A., Kassem, M., and Myklebost, O. (2010). Identification of target genes for wild type and truncated HMGA2 in mesenchymal stem-like cells. BMC Cancer *10*, 329.

Hill, D.A., Pedulla, M.L., and Reeves, R. (1999). Directional binding of HMG-I(Y) on four-way junction DNA and the molecular basis for competitive binding with HMG-1 and histone H1. Nucleic Acids Research *27*, 2135–2144.

Hiragami-Hamada, K., Soeroes, S., Nikolov, M., Wilkins, B., Kreuz, S., Chen, C., La Rosa-Vel aacute zquez, De, I.A., Zenn, H.M., Kost, N., Pohl, W., et al. (2016). Dynamic and flexible H3K9me3 bridging via HP1β dimerization establishes a plastic state of condensed chromatin. Nature Communications *7*, 1–16.

Hiratani, I., Ryba, T., Itoh, M., Yokochi, T., Schwaiger, M., Chang, C.-W., Lyou, Y., Townes, T.M., Schübeler, D., and Gilbert, D.M. (2008). Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol *6*, e245.

Hock, R., Furusawa, T., Ueda, T., and Bustin, M. (2007). HMG chromosomal proteins in development and disease. Trends in Cell Biology *17*, 72–79.

Holmberg, A., Blomstergren, A., Nord, O., Lukacs, M., Lundeberg, J., and Uhlen, M. (2005). The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. Electrophoresis *26*, 501–510.

Hoppe, P.S., Schwarzfischer, M., Loeffler, D., Kokkaliaris, K.D., Hilsenbeck, O., Moritz, N., Endele, M., Filipczyk, A., Gambardella, A., Ahmed, N., et al. (2016). Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. Nature *535*, 299–302.

Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D. (2006). The UCSC Known Genes. Bioinformatics *22*, 1036–1046.

Huth, J.R., Bewley, C.A., Nissen, M.S., Evans, J.N., Reeves, R., Gronenborn, A.M., and Clore, G.M. (1997). The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif. Nat. Struct. Biol. *4*, 657–665.

Ing-Simmons, E., Seitan, V.C., Faure, A.J., Flicek, P., Carroll, T., Dekker, J., Fisher, A.G., Lenhard, B., and Merkenschlager, M. (2015). Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. Genome Research *25*, 504–513.

Isakova, A., Berset, Y., Hatzimanikatis, V., and Deplancke, B. (2016). Quantification of Cooperativity in Heterodimer-DNA Binding Improves the Accuracy of Binding Specificity Models. J. Biol. Chem. *291*, 10293–10306.

Jain, D., Baldi, S., Zabel, A., Straub, T., and Becker, P.B. (2015). Active promoters give rise to false positive “Phantom Peaks” in ChIP-seq experiments. Nucleic Acids Research gkv637–10.

Jang, C.-W., Shibata, Y., Starmer, J., Yee, D., and Magnuson, T. (2015). Histone H3.3 maintains genome integrity during mammalian development. Genes & Development *29*, 1377–1392.

Jermann, P., Hoerner, L., Burger, L., and Schubeler, D. (2014). Short sequences can efficiently recruit histone H3 lysine 27 trimethylation in the absence of enhancer activity and DNA methylation. Proc Natl Acad Sci U S A *111*, E3415–E3421.

Johns, E.W. (1982). History, definitions and problems (The HMG chromosomal proteins).

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human

transcription factors. Cell *152*, 327–339.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384–388.

Joshi, A.A., and Struhl, K. (2005). Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation. Molecular Cell *20*, 971–978.

Joshi, S.R., Sarpong, Y.C., Peterson, R.C., and Scovell, W.M. (2012). Nucleosome dynamics: HMGB1 relaxes canonical nucleosome structure to facilitate estrogen receptor binding. Nucleic Acids Research *40*, 10161–10171.

Jung, Y., and Lippard, S.J. (2003). Nature of Full-Length HMGB1 Binding to Cisplatin-Modified DNA †. Biochemistry *42*, 2664–2671.

Kang, E., Wu, G., Ma, H., Li, Y., Tippner-Hedges, R., Tachibana, M., Sparman, M., Wolf, D.P., Schöler, H.R., and Mitalipov, S. (2014). Nuclear reprogramming by interphase cytoplasm of two-cell mouse embryos. Nature *509*, 101–104.

Kaplan, T., Li, X.-Y., Sabo, P.J., Thomas, S., Stamatoyannopoulos, J.A., Biggin, M.D., and Eisen, M.B. (2011). Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early Drosophila Development. PLoS Genet *7*, e1001290.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Research *12*, 996–1006.

Kidder, B.L., Hu, G., and Zhao, K. (2011). ChIP-Seq: technical considerations for obtaining high-quality data. Nat Immunol *12*, 918–922.

Kilpinen, H., Waszak, S.M., Gschwind, A.R., Raghav, S.K., Witwicki, R.M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N.I., et al. (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science *342*, 744–747.

Kim, J., Cantor, A.B., Orkin, S.H., and Wang, J. (2009). Use of in vivo biotinylation to study protein-protein and protein-DNA interactions in mouse embryonic stem cells. Nature Protocols *4*, 506–517.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. Cell *132*, 1049–1061.

Krebs, A.R., Dessus-Babus, S., Burger, L., and Schübeler, D. (2014). High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. Elife *3*, e04094.

Kugler, J.E., Deng, T., and Bustin, M. (2012). The HMGN family of chromatin-

binding proteins: dynamic modulators of epigenetic processes. Biochim. Biophys. Acta *1819*, 652–656.

Kuznetsov, V.A., Singh, O., and Jenjaroenpun, P. (2010). Statistics of protein-DNA binding and the total number of binding sites for a transcription factor in the mammalian genome. BMC Genomics *11*, S12–S27.

Lalmansingh, A.S., Karmakar, S., Jin, Y., and Nagaich, A.K. (2012). Multiple modes of chromatin remodeling by Forkhead box proteins. BBA - Gene Regulatory Mechanisms *1819*, 707–715.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology *10*, R25.

Langowski, T.E.A.J. (2015). The effect of DNA supercoiling on nucleosome structure and stability. Journal of Physics: Condensed Matter *27*, 064105.

Laugesen, A., and Helin, K. (2014). Chromatin Repressive Complexes in Stem Cells, Development, and Cancer. Stem Cell *14*, 735–751.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology *15*, R29.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol *9*, e1003118.

Lee, J.H. (2001). Identification and Characterization of the DNA Binding Domain of CpG-binding Protein. Journal of Biological Chemistry *276*, 44669–44676.

Lee, K.-B., and Thomas, J.O. (2000). The effect of the acidic tail on the DNA-binding properties of the HMG1,2 class of proteins: insights from tail switching and tail removal. Journal of Molecular Biology *304*, 135–149.

Lehnertz, B., Ueda, Y., Derijck, A., and Braunschweig, U. (2003). Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Current Biology *13*, 1192–1200.

Levine, M. (2010). Transcriptional Enhancers in Animal Development and Evolution. Current Biology *20*, R754–R763.

Lia, G., Bensimon, D., Croquette, V., Allemand, J.-F., Dunlap, D., Lewis, D.E.A., Adhya, S., and Finzi, L. (2003). Supercoiling and denaturation in Gal repressor/heat unstable nucleoid protein (HU)-mediated DNA looping. Proc Natl Acad Sci U S A *100*, 11373–11377.

Lienert, F., Mohn, F., Tiwari, V.K., Baubec, T., Roloff, T.C., Gaidatzis, D., Stadler, M.B., and Schübeler, D. (2011). Genomic prevalence of heterochromatic H3K9me2 and transcription do not discriminate pluripotent from terminally

differentiated cells. PLoS Genet *7*, e1002090.

Lippard, S.J., Ohndorf, U.-M., Rould, M.A., He, Q., and Pabo, C.O. (1999). Basis for recognition of cisplatin-modified DNA by high-mobility-group proteins. Nature *399*, 708–712.

Luger, K. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251–260.

Lund, T., Holtlund, J., and Laland, S.G. (1985). On the phosphorylation of low molecular mass HMG (high mobility group) proteins in Ehrlich ascites cells. FEBS Letters.

Martens, J.H.A., O'Sullivan, R.J., Braunschweig, U., Opravil, S., Radolf, M., Steinlein, P., and Jenuwein, T. (2005). The profile of repeat-associated histone lysine methylation states in the mouse epigenome. The EMBO Journal *24*, 800–812.

Mathelier, A., and Wasserman, W.W. (2013). The next generation of transcription factor binding site prediction. PLoS Comput Biol *9*, e1003214.

Mattout, A., Cabianca, D.S., and Gasser, S.M. (2015). Chromatin states and nuclear organization in development — a view from the nuclear lamina. Genome Biology 1–15.

Maurano, M., Wang, H., John, S., Shafer, A., Canfield, T., Lee, K., and Stamatoyannopoulos, J.A. (2014). DNA methylation modulates transcription factor occupancy chiefly at sites of high intrinsic cell-type variability.

McKay, D.J., Klusza, S., Penke, T.J.R., Meers, M.P., Curry, K.P., McDaniel, S.L., Malek, P.Y., Cooper, S.W., Tatomer, D.C., Lieb, J.D., et al. (2015). Interrogating the function of metazoan histones using engineered gene clusters. Developmental Cell *32*, 373–386.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells. PLoS Genet *6*, e1001244–10.

Merika, M., and Thanos, D. (2001). Enhanceosomes. Current Opinion in Genetics & Development *11*, 205–208.

Merkenschlager, M., and Odom, D.T. (2013). CTCF and cohesin: linking gene regulatory elements with their targets. Cell *152*, 1285–1297.

Meuleman, W., Yen, A., Heravi-Moussavi, A., Wang, J., Amin, V., Sarkar, A., Quon, G., Wu, Y.-C., Pfenning, A., Wang, X., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.

Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M., and Schübeler, D. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Molecular Cell *30*, 755–766.

Moris, N., Pina, C., and Arias, A.M. (2016). Transition states and cell fate decisions in epigenetic landscapes. Nature Reviews Genetics *17*, 693–703.

Morishita, A., Zaidi, M.R., Mitoro, A., Sankarasharma, D., Szabolcs, M., Okada, Y., D'Armiento, J., and Chada, K. (2013). HMGA2 Is a Driver of Tumor Metastasis. Cancer Research *73*, 4289–4299.

Naughton, C., Avlonitis, N., Corless, S., Prendergast, J.G., Mati, I.K., Eijk, P.P., Cockroft, S.L., Bradley, M., Ylstra, B., and Gilbert, N. (2013). Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. Nature Structural &Amp; Molecular Biology *20*, 387–395.

Neph, S., Stergachis, A.B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J.A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. Cell *150*, 1274–1286.

Odom, D.T. (2011). Identification of Transcription Factor-DNA Interactions In Vivo. Subcell. Biochem. *52*, 175–191.

Ohndorf, U.M., Rould, M.A., He, Q., Pabo, C.O., and Lippard, S.J. (1999). Basis for recognition of cisplatin-modified DNA by high-mobility-group proteins. Nature *399*, 708–712.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nature Reviews Genetics 1–13.

Ooi, L., and Wood, I.C. (2007). Chromatin crosstalk in development and disease: lessons from REST. Nature Reviews Genetics *8*, 544–554.

Pallier, C., Scaffidi, P., Chopineau-Proust, S., Agresti, A., Nordmann, P., Bianchi, M.E., and Marechal, V. (2003). Association of chromatin proteins high mobility group box (HMGB) 1 and HMGB2 with mitotic chromosomes. Mol. Biol. Cell *14*, 3414–3426.

Papantonis, A., and Cook, P.R. (2013). Transcription Factories: Genome Organization and Gene Regulation. Chem. Rev. *113*, 8683–8705.

Pasheva, E.A., Pashev, I.G., and Favre, A. (1998). Preferential binding of high mobility group 1 protein to UV-damaged DNA. Role of the COOH-terminal domain. J. Biol. Chem. *273*, 24730–24736.

Peter, S., Yu, H., Ivanyi-Nagy, R., and Dröge, P. (2016). Cell-based high-throughput compound screening reveals functional interaction between oncofetal HMGA2 and topoisomerase I. Nucleic Acids Research gkw759–10.

Phair, R.D., Scaffidi, P., Elbi, C., Vecerová, J., Dey, A., Ozato, K., Brown, D.T., Hager,

G., Bustin, M., and Misteli, T. (2004). Global nature of dynamic protein-chromatin interactions in vivo: three-dimensional genome scanning and dynamic interaction networks of chromatin proteins. Molecular and Cellular Biology *24*, 6393–6402.

Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W., and Sandelin, A. (2009). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Research *38*, gkp950–D110.

Prescott, S.L., Srinivasan, R., Marchetto, M.C., Grishina, I., Narvaiza, I., Selleri, L., Gage, F.H., Swigut, T., and Wysocka, J. (2015). Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. Cell *163*, 68–83.

Radic, M.Z., Saghbini, M., Elton, T.S., Reeves, R., and Hamkalo, B.A. (1992). Hoechst 33258, distamycin A, and high mobility group protein I (HMG-I) compete for binding to mouse satellite DNA. Chromosoma *101*, 602–608.

Raveh-Sadka, T., Levo, M., Shabi, U., Shany, B., Keren, L., Lotan-Pompan, M., Zeevi, D., Sharon, E., Weinberger, A., and Segal, E. (2012). Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nature Genetics *44*, 743–750.

Redmond, A.M., Byrne, C., Bane, F.T., Brown, G.D., Tibbitts, P., Brien, K.O.R., Hill, A.D.K., Carroll, J.S., and Young, L.S. (2014). Genomic interaction between ER and HMGB2 identifies DDX18 as a novel driver of endocrine resistance in breast cancer cells. Oncogene *34*, 3871–3880.

Reeves, R. (2000). Structure and function of the HMGI (Y) family of architectural transcription factors. Environmental Health Perspectives *108*, 803.

Reeves, R., and Beckerbauer, L. (2001). HMGI/Y proteins: flexible regulators of transcription and chromatin structure. … Et Biophysica Acta (BBA)-Gene Structure and … *1519*, 13–29.

Reeves, R., and Nissen, M.S. (1990). The A.T-DNA-binding domain of mammalian high mobility group I chromosomal proteins. A novel peptide motif for recognizing DNA structure. J. Biol. Chem. *265*, 8573–8582.

Reeves, R., and Nissen, M.S. (1993). Interaction of high mobility group-I (Y) nonhistone proteins with nucleosome core particles. J. Biol. Chem. *268*, 21137–21146.

Reeves, R., and Wolffe, A.P. (1996). Substrate structure influences binding of the non-histone protein HMG-I(Y) to free nucleosomal DNA. Biochemistry *35*, 5063–5074.

Reeves, R., Leonard, W.J., and Nissen, M.S. (2000). Binding of HMG-I(Y) imparts architectural specificity to a positioned nucleosome on the promoter of the human interleukin-2 receptor alpha gene. Molecular and Cellular Biology *20*, 4666–4679.

Reeves, R. (2010). Nuclear functions of the HMG proteins. Biochim. Biophys. Acta *1799*, 3–14.

Ringnér, M. (2008). What is principal component analysis? Nature Biotechnology.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research *43*, e47–e47.

Rockowitz, S., Lien, W.-H., Pedrosa, E., Wei, G., Lin, M., Zhao, K., Lachman, H.M., Fuchs, E., and Zheng, D. (2014). Comparison of REST Cistromes across Human Cell Types Reveals Common and Context-Specific Functions. PLoS Comput Biol *10*, e1003671–17.

Roemer, S.C., Adelman, J., Churchill, M.E.A., and Edwards, D.P. (2008). Mechanism of high-mobility group protein B enhancement of progesterone receptor sequence-specific DNA binding. Nucleic Acids Research *36*, 3655–3666.

Roh, T.-Y., Cuddapah, S., and Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes & Development *19*, 542–552.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. Annu. Rev. Biochem. *79*, 233–269.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248–1253.

Ross-Innes, C.S., Brown, G.D., and Carroll, J.S. (2011). A co-ordinated interaction between CTCF and ER in breast cancer cells. BMC Genomics *12*, 593.

Sainsbury, S., Niesser, J., and Cramer, P. (2012). Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. Nature 1–5.

Sanchez-Giraldo, R., Acosta-Reyes, F.J., Malarkey, C.S., Saperas, N., Churchill, M.E.A., and Campos, J.L. (2015). Two high-mobility group box domains act together to underwind and kink DNA. Acta Cryst (2015). D71, 1423-1432 [Doi:10.1107/S1399004715007452] 1–10.

Sarkar, D., Gentleman, R., Lawrence, M., and Yao, Z. (2013). chipseq: A package for analyzing chipseq data (R package).

Scaffidi, P., and Bianchi, M.E. (2001). Spatially precise DNA bending is an essential activity of the sox2 transcription factor. J. Biol. Chem. *276*, 47296–47302.

Scaffidi, P., Misteli, T., and Bianchi, M.E. (2002). Release of chromatin protein HMGB1 by necrotic cells triggers inflammation. Nature *418*, 191–195.

Schübeler, D. (2015). Function and information content of DNA methylation.

Nature *517*, 321–326.

Sebé-Pedrós, A., Ballaré, C., Parra-Acero, H., Chiva, C., Tena, J.J., Sabidó, E., Gómez-Skarmeta, J.L., Di Croce, L., and Ruiz-Trillo, I. (2016). The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity. Cell *165*, 1224–1237.

Segal, E., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.

Sgarra, R., Zammitti, S., Sardo, Lo, A., Maurizio, E., Arnoldo, L., Pegoraro, S., Giancotti, V., and Manfioletti, G. (2010). HMGA molecular network: From transcriptional regulation to chromatin remodeling. Biochimica Et Biophysica Acta (BBA) - Gene Regulatory Mechanisms *1799*, 37–47.

Shachar, S., Voss, T.C., Pegoraro, G., Sciascia, N., and Misteli, T. (2015). Identification of Gene Positioning Factors Using High-Throughput Imaging Mapping. Cell *162*, 911–923.

Shah, M.A., Denton, E.L., Arrowsmith, C.H., Lupien, M., and Schapira, M. (2014). A global assessment of cancer genomic alterations in epigenetic mechanisms. Epigenetics & Chromatin *7*, 29–15.

Sheflin, L.G., Fucile, N.W., and Spaulding, S.W. (1993). The specific interactions of HMG 1 and 2 with negatively supercoiled DNA are modulated by their acidic C-terminal domains and involve cysteine residues in their HMG 1/2 boxes. Biochemistry *32*, 3238–3248.

Simicevic, J., Schmid, A.W., Gilardoni, P.A., Zoller, B., Raghav, S.K., Krier, I., Gubelmann, C., Lisacek, F., Naef, F., Moniatte, M., et al. (2013). Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics. Nat Meth *10*, 570–576.

Singh, I., Ozturk, N., Cordero, J., Mehta, A., Hasan, D., Cosentino, C., Sebastian, C., Krüger, M., Looso, M., Carraro, G., et al. (2015). High mobility group protein-mediated transcription requires DNA damage marker γ-H2AX. Nature Publishing Group *25*, 837–850.

Singh, M., D'Silva, L., and Holak, T.A. (2006). DNA-binding properties of the recombinant high-mobility-group-like AT-hook-containing region from human BRG1 protein. Biol. Chem. *387*, 1469–1478.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends in Biochemical Sciences *39*, 381–399.

Smolle, M., and Workman, J.L. (2013). Transcription-associated histone modifications and cryptic transcription. BBA - Gene Regulatory Mechanisms *1829*, 84–97.

Soler, E., Andrieu-Soler, C., de Boer, E., Bryne, J.C., Thongjuea, S., Stadhouders, R.,

Palstra, R.J., Stevens, M., Kockx, C., van IJcken, W., et al. (2010). The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. Genes & Development *24*, 277–289.

Soler, E., Andrieu-Soler, C., de Boer, E., Bryne, J.C., Thongjuea, S., Rijkers, E., Demmers, J., van IJcken, W., and Grosveld, F. (2011). A systems approach to analyze transcription factors in mammalian cells. Methods *53*, 151–162.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2015). Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. Cell *161*, 1–27.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nature Reviews Genetics *13*, 613–626.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Stott, K., Watson, M., Bostock, M.J., Mortensen, S.A., Travers, A., Grasser, K.D., and Thomas, J.O. (2014). Structural insights into the mechanism of negative regulation of single HMG-box proteins by the acidic tail domain. Journal of Biological Chemistry jbc.M114.591115.

Strauss, F., and Varshavsky, A. (1984). A protein binds to a satellite DNA repeat at three specific sites that would be brought into mutual proximity by DNA folding in the nucleosome. Cell *37*, 889–901.

Strichman-Almashanu, L.Z., Bustin, M., and Landsman, D. (2003). Retroposed copies of the HMG genes: a window to genome dynamics. Genome Research *13*, 800–812.

Stros, M. (2010). HMGB proteins: interactions with DNA and chromatin. Biochim. Biophys. Acta *1799*, 101–113.

Sun, M., Song, C.X., Huang, H., Frankenberger, C.A., Sankarasharma, D., Gomes, S., Chen, P., Chen, J., Chada, K.K., He, C., et al. (2013). HMGA2/TET1/HOXA9 signaling pathway regulates breast cancer growth and metastasis. Proceedings of the National Academy of Sciences *110*, 9920–9925.

Takahashi, K., and Yamanaka, S. (2016). A decade of transcription factor-mediated reprogramming to pluripotency. Nat. Rev. Mol. Cell Biol. 1–11.

Teytelman, L., Thurtle, D.M., and Rine, J. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins.

Thåström, A., Lowary, P.T., Widlund, H.R., Cao, H., Kubista, M., and Widom, J. (1999). Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. Journal of Molecular Biology *288*, 213–229.

Thomas, J.O., and Travers, A.A. (2001). HMG1 and 2, and related "architectural" DNA-binding proteins. Trends in Biochemical Sciences *26*, 167–174.

Tootle, T.L., and Rebay, I. (2005). Post-translational modifications influence transcription factor activity: a view from the ETS superfamily. Bioessays *27*, 285–298.

Travers, A.A. (2003). Priming the nucleosome: a role for HMGB proteins? EMBO Rep *4*, 131–136.

Tropberger, P., Pott, S., Keller, C., Kamieniarz-Gdula, K., Caron, M., Richter, F., Li, G., Mittler, G., Liu, E.T., Bühler, M., et al. (2013). Regulation of Transcription through Acetylation of H3K122 on the Lateral Surface of the Histone Octamer. Cell *152*, 859–872.

Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Meth *5*, 829–834.

Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. Nature *474*, 516–520.

van Steensel, B. (2011). Chromatin: constructing the big picture. The EMBO Journal *30*, 1885–1895.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nature Reviews Genetics *10*, 252–263.

Verrijdt, G., Haelens, A., Schoenmakers, E., Rombauts, W., and Claessens, F. (2002). Comparative analysis of the influence of the high-mobility group box 1 protein on DNA binding and transcriptional activation by the androgen, glucocorticoid, progesterone and mineralocorticoid receptors. Biochem. J. *361*, 97–103.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer Evolution across 20 Mammalian Species. Cell *160*, 554–566.

Villar, D., Flicek, P., and Odom, D.T. (2014). Evolution of transcription factor binding in metazoans — mechanisms and functional implications. Nature Reviews Genetics 1–13.

Vlijm, R., Lee, M., Lipfert, J., Lusser, A., Dekker, C., and Dekker, N.H. (2015). Nucleosome Assembly Dynamics Involve Spontaneous Fluctuations in the Handedness of Tetrasomes. CellReports *10*, 216–225.

Voss, T.C., and Hager, G.L. (2013). Dynamic regulation of transcriptional states by chromatin and transcription factors. Nature Publishing Group *15*, 69–81.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Research *22*, 1798–1812.

Wang, J., Rao, S., Chu, J., Shen, X., Levasseur, D.N., Theunissen, T.W., and Orkin, S.H. (2006). A protein interaction network for pluripotency of embryonic stem cells. Nature *444*, 364–368.

Wang, Q., Zeng, M., Wang, W., and Tang, J. (2007). The HMGB1 acidic tail regulates HMGB1 DNA binding specificity by a unique mechanism. Biochemical and Biophysical Research ….

Watanabe, S., Radman-Livaja, M., Rando, O.J., and Peterson, C.L. (2013). A histone acetylation switch regulates H2A.Z deposition by the SWR-C remodeling enzyme. Science *340*, 195–199.

Watson, M., Stott, K., Fischl, H., Cato, L., and Thomas, J.O. (2014). Characterization of the interaction between HMGB1 and H3--a possible means of positioning HMGB1 in chromatin. Nucleic Acids Research *42*, 848–859.

Whalen, S., Truty, R.M., and Pollard, K.S. (2016). Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. Nature Genetics *48*, 488–496.

Whitington, T., Frith, M.C., Johnson, J., and Bailey, T.L. (2011). Inferring transcription factor complexes from ChIP-seq data. Nucleic Acids Research *39*, e98–e98.

Wilson, W.D., Tanious, F.A., Barton, H.J., Jones, R.L., Fox, K., Wydra, R.L., and Strekowski, L. (1990). DNA sequence dependent binding modes of 4',6-diamidino-2-phenylindole (DAPI). Biochemistry *29*, 8452–8461.

Winter, N., Nimzyk, R., Bösche, C., Meyer, A., and Bullerdiek, J. (2011). Chromatin Immunoprecipitation to Analyze DNA Binding Sites of HMGA2. PLoS ONE *6*, e18837–e18838.

Wittkopp, P.J., and Kalay, G. (2011). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics *13*, 59–69.

Wood, L.J., Mukherjee, M., Dolde, C.E., Xu, Y., Maher, J.F., Bunton, T.E., Williams, J.B., and Resar, L.M. (2000). HMG-I/Y, a new c-Myc target gene and potential oncogene. Molecular and Cellular Biology *20*, 5490–5502.

Woodcock, C.L., Skoultchi, A.I., and Fan, Y. (2006). Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. Chromosome Res *14*, 17–25.

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics *8*, 206–216.

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C.L., Haase, J., Janes, J., Huss, J.W., et al. (2009). BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biology *10*, R130.

Xu, M., Sharma, P., Pan, S., Malik, S., Roeder, R.G., and Martinez, E. (2011). Core promoter-selective function of HMGA1 and Mediator in Initiator-dependent transcription. Genes & Development *25*, 2513–2524.

Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M., et al. (2013). Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites. Cell *154*, 801–813.

Yáñez-Cuna, J.O., Kvon, E.Z., and Stark, A. (2013). Deciphering the transcriptional cis-regulatory code. Trends in Genetics *29*, 11–22.

Ye, Z., Chen, Z., Sunkel, B., Frietze, S., Huang, T.H.M., Wang, Q., and Jin, V.X. (2016). Genome-wide analysis reveals positional-nucleosome-oriented binding pattern of pioneer factor FOXA1. Nucleic Acids Research *44*, 7540–7554.

Yie, J., Merika, M., Munshi, N., Chen, G., and Thanos, D. (1999). The role of HMG I(Y) in the assembly and function of the IFN-beta enhanceosome. The EMBO Journal *18*, 3074–3089.

Youn, J.H., and Shin, J.S. (2006). Nucleocytoplasmic Shuttling of HMGB1 Is Regulated by Phosphorylation That Redirects It toward Secretion. The Journal of Immunology *177*, 7889–7897.

Yumoto, Y., Shirakawa, H., Yoshida, M., Suwa, A., Watanabe, F., and Teraoka, H. (1998). High Mobility Group Proteins 1 and 2 Can Function as DNA-Binding Regulatory Components for DNA-Dependent Protein Kinase In Vitro. Journal of Biochemistry *124*, 519–527.

Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: establishing competence for gene expression. Genes & Development *25*, 2227–2241.

Zha, L., Wang, Z., Tang, W., Zhang, N., Liao, G., and Huang, Z. (2012). Genome-wide analysis of HMGA2 transcription factor binding sites by ChIP on chip in gastric carcinoma cells. Mol Cell Biochem *364*, 243–251.

Zhang, Q., and Wang, Y. (2008). High mobility group proteins and their post-translational modifications. Biochimica Et Biophysica Acta (BBA) - Proteins and Proteomics *1784*, 1159–1166.

Zhang, R., Chen, W., and Adams, P.D. (2007). Molecular dissection of formation of senescence-associated heterochromatin foci. Molecular and Cellular Biology *27*, 2343–2358.

Zhang, S., Zhu, I., Deng, T., Furusawa, T., Rochman, M., Vacchio, M.S., Bosselut, R., Yamane, A., Casellas, R., Landsman, D., et al. (2016). HMGN proteins modulate

chromatin regulatory sites and gene expression during activation of naïve B cells. Nucleic Acids Research gkw323–15.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biology *9*, R137.

Zhao, K., Käs, E., Gonzalez, E., and Laemmli, U.K. (1993). SAR-dependent mobilization of histone H1 by HMG-I/Y in vitro: HMG-I/Y is enriched in H1-depleted chromatin. The EMBO Journal *12*, 3237–3247.

Zhao, R., Deibler, R.W., Lerou, P.H., Ballabeni, A., Heffner, G.C., Cahan, P., Unternaehrer, J.J., Kirschner, M.W., and Daley, G.Q. (2014). A nontranscriptional role for Oct4 in the regulation of mitotic entry. Proceedings of the National Academy of Sciences 201417518.

Zhuang, Q., Smallman, H., Lambert, S.J., and Sodngam, S.S. (2014). Cofractionation of HMGB proteins with histone dimers. Analytical ….

(2013). Epigenetics: Development and Disease (Dordrecht: Springer Netherlands).

11 Acknowledgements

Thanks to Dirk Schübeler for giving me the opportunity to confront me with this challenging project. Thanks to Lukas Burger for all the exciting moments in the data analysis and fruitful discussion.

Thanks to Mildred Alejandra Gutierrez Herrera, my soul mate that accompanied me in the last 2 and half years daily.

Thanks to my family with whom I shared the pains of being far away and the joy of becoming independent.

Thanks to my friends and my colleagues for the good time passed together, especially when experiments were not working.